International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-025-00816-w

REGULAR PAPER

Synthetic Data Generation for Healthcare: Exploring Generative Adversarial Networks Variants for Medical Tabular Data

Halal Abdulrahman Ahmed[1] · Juan A. Nepomuceno[1] · Belén Vega-Márquez[1] · Isabel A. Nepomuceno-Chamorro[1]

Received: 7 October 2024 / Accepted: 8 May 2025

Abstract
Recently, the medical and healthcare fields have experienced significant improvements. However, the restrictions of ethical constraints, privacy regulations, and preservation for sharing sensitive personal information limit access to real patient data. Synthetic datasets with generative models are considered one of the most reliable solutions that meet strict data protection requirements. Synthetic data are created in a controlled environment but possess the same statistical and structural properties as real data. In this work, we generate synthetic data using six variations of generative adversarial networks (GANs): GAN, CGAN, CTGAN, CRAMER GAN, DRAGAN, and WGAN. We explore the efficacy of synthetic data in three distinct healthcare datasets: Breast Cancer Wisconsin (Diagnostic), Lung Cancer Patient, and Fetal Cardiotocography CTG. To evaluate the performance of these generated datasets in classification tasks, we employ two diverse classifiers, namely XGBoost and SVM. In addition, we employ correlation and statistical analyses to scrutinise GAN models, identifying optimal variants for specific data generation tasks. Our experimental framework encompasses the examination of original (real), synthetic, and hybrid (original and synthetic) datasets. Our findings highlight a notable improvement in classification accuracy when using advanced GAN models such as CGAN and CTGAN to generate tabular data. This research sheds light on the potential of synthetic data in bolstering data privacy while facilitating meaningful insights in the realm of healthcare analytics.

Keywords Synthetic data generation · Generative adversarial networks · Privacy-preserving data · Deep learning

1 Introduction

Over the past decade, the medical field has seen remarkable advancements due to increased data availability, machine learning techniques, and artificial intelligence. High-quality datasets are essential for training and testing machine learning models in healthcare, but strict privacy restrictions, data scarcity, and ethical constraints limit access to such data [1, 2]. To address this challenge, researchers have developed techniques such as Variational Autoencoders (VAE) [3] and Generative Adversarial Networks (GAN) to generate synthetic data. Synthetic data, created through algorithms, is critical when real data is challenging, expensive, or limited to acquire. It preserves the statistical and structural characteristics of real data while maintaining patient privacy. In healthcare, synthetic data helps develop accurate machine learning models, improve treatments, and understand diseases [4, 5], enhancing the quality of healthcare and biometric [6] research.

GANs, in particular, generate synthetic data resembling real medical records without containing actual patient details, thus overcoming data scarcity [7–9].

GANs have shown remarkable results in generating realistic images and videos for various applications. Han et al. [10] successfully generated brain MR images. Jin et al. [11] focused on generating CT images encompassing both nodules and adjacent tissues. Bhagat et al. [12] generated chest X-ray images of pneumonia patients. Uzunova et al. [13] generated realistic and high-resolution 2D and 3D medical data. Munia et al. [14] generated synthetic electrocardiogram (ECG) data. Moreover, GANs are also used in other types of data; for instance, Lei Xu et al. presented Conditional Tabular GAN (CTGAN) [15], which aims to generate high-quality tabular datasets encompassing various data types. CTGAN employs a conditional generator and an innovative training-by-sampling approach to address the generation of imbalanced data. The CTGAN model adopts a mode-specific

✉ Halal Abdulrahman Ahmed
halahmabd@alum.us.es

[1] Department of Computer Languages and Systems, University of Seville, Seville 41012, Spain

Published online: 24 May 2025

normalisation technique to effectively handle the complexity of generating multi-modal numerical columns. Li et al. [16] proposed EHR-M-GAN to generate mixed-type time-series EHR data. Mottini et al. [17] employed CRAMER GAN to generate synthetic Personal Name Records (PNR) by training the model on real PNRs with numerical and categorical features. The experimental results indicate that the generative model produces realistic synthetic data that closely matches the distribution of real PNRs. Azman et al. [18] used DRAGAN to generate synthetic medical images; these images are subsequently employed to classify lung lesions into benign and malignant categories using ShuffleNet. Chin-Cheong et al. [19] employed WGAN to generate high-quality numerical and categorical heterogeneous EHR data in terms of both data fidelity and data utility by combining it with differential privacy (DP). Hussain et al. [20] presented a solution for addressing the data deficiency in COVID-19 chest X-ray images by using the Wasserstein Generative Adversarial Network (WGAN). The experiments demonstrated WGAN's effectiveness in generating synthetic X-ray images. Several research studies have utilised WGAN [21–23].

Studies have explored various GAN-based models for generating synthetic medical tabular data that maintain statistical characteristics and model compatibility of original data [24–26]. Although the generation of synthetic data is rapidly increasing, the generation of synthetic tabular data still presents unique challenges compared to other data types [27]. Relatively, there are less studies on the generation of synthetic tabular data compared to other types of data, such as images, text or speech [28–30].

This study makes several significant contributions to the field of synthetic data generation in healthcare. Our primary contribution lies in comparing six different GAN variants - GAN, CGAN, CTGAN, CRAMER GAN, DRAGAN, and WGAN- to examine their functionalities and efficacy in generating high-quality synthetic tabular data for medical problems. In pursuit of our objective, we selected three different publicly available medical datasets, including Breast Cancer Wisconsin (Diagnostic) [31], Lung Cancer Patient [32], and Fetal Cardiotocography (CTG) [33].

To access the similarity and usability of synthetic medical data and real patient data, we use two widely recognised classification algorithms, including XGBoost and Support Vector Machine (SVM). We will employ two statistical methods to analyze the relationships between variables in the datasets (Pearson and Spearman correlations). Furthermore, to evaluate the performance of the GAN models and making sure the results are accurate and reliable, we will conduct statistical evaluations by using the Statistical Tests for Algorithms Comparison (STAC) platform.

The remainder of this article is organised as follows: Section 2 describes the methods for generating synthetic data; Section 3 describes the materials used; Section 4 reports and discusses the results obtained; and Section 5 presents the conclusions and potential future work.

2 Methods

This section delves into the methods used to generate synthetic medical data using Generative Adversarial Networks (GAN) methods. We provide detailed explanations of the selected GAN variants and classifiers used to measure the performance of the synthetic data. We have decided to check whether the behaviour of classifying the generated data is similar to the behaviour of real data. This aims to ensure the validity of the generated data through the performance of the classifiers and the understanding of the limitations of the GAN models.

2.1 Generative Adversarial Networks (GANs)

GAN is an artificial intelligence model introduced by Ian Goodfellow [34]. It is a robust framework for training generative models, capable of generating new data that resembles a given dataset. GANs are composed of two sub-networks: the generator (G) and the discriminator (D), based on the idea of zero-sum games [35]. These components are two neural networks simultaneously trained through a competitive procedure in which one network attempts to gain an advantage while the other network experiences an equivalent loss. **Generator(G):** The main task of the generator is to generate synthetic data that resembles the real data of the training set. It takes random noise as input and transforms it into synthetic data. **Discriminator(D):** The discriminator is a binary classifier designed to differentiate between real data from the training set and synthetic data generated by the generator. It receives both real and generated data samples as input and returns a probability score indicating the likelihood of the input being real.

These two sub-networks compete to improve their predictive accuracy by pitting themselves against one another. When the generator is training the discriminator is inactive, and when the discriminator is training, the generator is inactive. The generator becomes better at generating higher-quality synthetic data through this competitive process, so the discriminator becomes better at flagging artificially generated data. In Figure 1, the architecture of GAN can be observed.

2.1.1 Conditional GAN (CGAN)

Conditional GAN [36] is an extension of the standard GAN model. In a standard GAN, the generator learns to generate synthetic data from random noise, whereas the discriminator attempts to discriminate between real and fake samples.

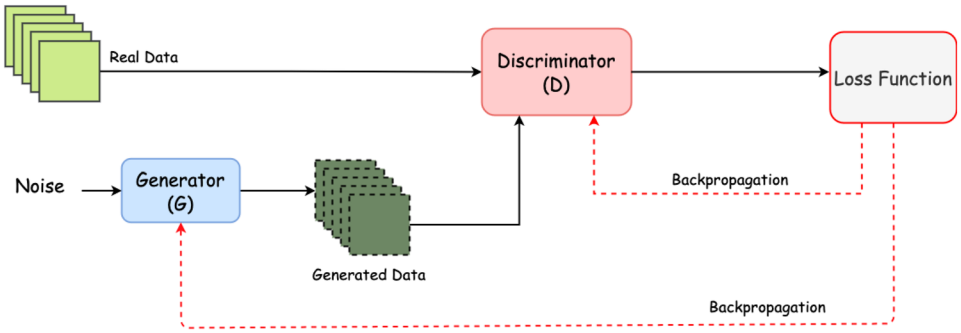


Fig. 1 Illustration of the GAN architecture

However, CGAN adds additional conditional information or conditional labels to the data generator and the discriminator for more targeted and controlled data generation. This conditional information instructs the generator to generate samples that meet particular criteria or belong to a specific class. The discriminator considers this additional condition when discriminating between real and fake data. Typically, conditional data is provided as an additional input to neural networks. The conditional GAN architecture provides more targeted and controlled data generation [37].

2.1.2 Conditional Tabular GAN (CTGAN)

CTGAN [15] is a GAN-based model developed particularly to generate synthetic tabular data with conditional attributes. Similar to CGAN, in CTGAN, the generator inputs random noise and conditional information to generate synthetic tabular data that adheres to the specified conditions. CTGAN can generate new rows of tabular data that satisfy given constraints or adhere to specific criteria. CTGAN provides a powerful tool for generating conditional tabular data and allows users to control and influence the characteristics of the data generated through conditional inputs. The architecture of GTGAN is adapted to handle the constraints and dependencies found in tabular datasets.

2.1.3 CRAMER GAN

CRAMER GAN [38] uses the Cramer distance, a more stable metric, to address training instabilities and mode collapse issues faced by standard GANs. Cramer distance provides a computationally tractable measure of the discrepancy between probability distributions. By using the Cramer distance, CRAMER GAN improves the quality of generated samples and encourages diversity in the generated data distribution, resulting in a more faithful representation of the true data distribution. CramerGAN is an example of how different distance metrics can be used in GANs to achieve specific objectives and overcome training challenges.

2.1.4 Deep Regret Analytic GAN (DRAGAN)

DRAGAN [39] is a proposed regularisation technique to train GANs, address the issue of mode collapse [40] and improve training stability. Mode collapse occurs when the GAN's generator fails to capture all the modes or diverse patterns in the real data distribution, resulting in a lack of diversity in the generated samples. DRAGAN proposes a novel approach to mitigate this problem by constraining the discriminator gradients around real data points. By employing the regularisation technique of DRAGAN, the training process of GANs achieves stability and a reduced susceptibility to mode collapse. t enables faster training, improved stability, and better modelling performance compared to other stable training procedures like WGAN-GP (Wasserstein GAN with Gradient Penalty) [41].

2.1.5 Wasserstein GAN (WGAN)

Wasserstein GAN [42] is an advanced variant of GANs. It utilises the Wasserstein distance as the loss function, offering a more meaningful measure of data distribution dissimilarity compared to traditional GANs. WGAN replaces the binary discriminator with a critic and enforces a 1-Lipschitz constraint for stability during training. As a result, WGAN achieves more stable training, reduces mode collapse issues [40], and provides a better evaluation of the performance of the generator. Its effectiveness has made WGAN popular in generative models and deep learning research.

2.2 Classifiers

2.2.1 Extreme Gradient Boosting Classifier

Extreme Gradient Boosting (XGBoost) [43] classifier is a machine learning algorithm that implements gradient-boosted decision trees, excelling in classification and regression tasks. Furthermore, the boosting technique employed in this approach is regularised, allowing for the automated handling of missing values. Additionally, the algorithm has been specifically developed to show high efficiency, flexi-

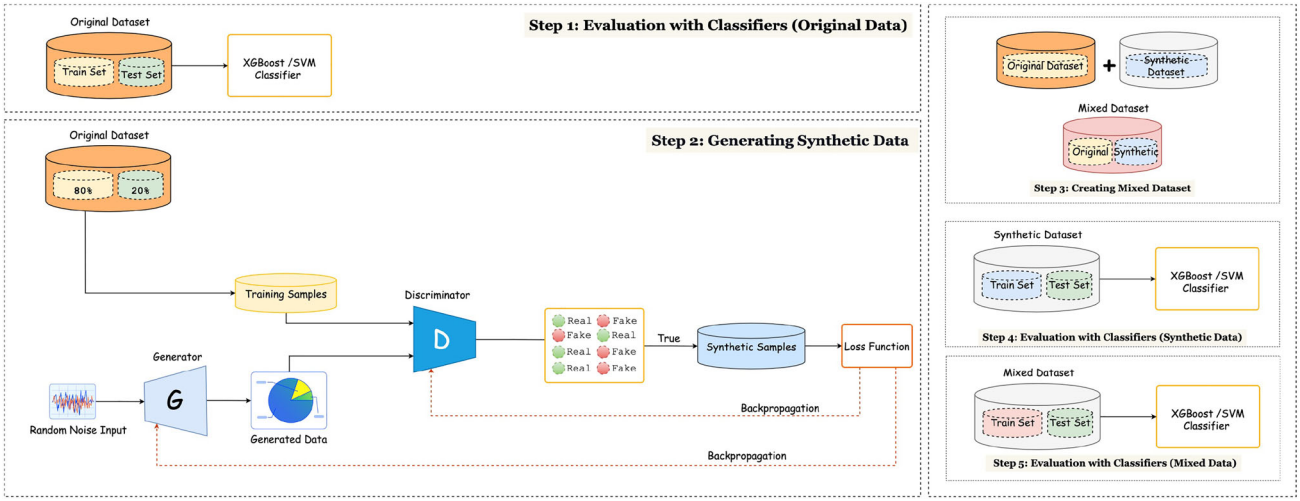


Fig. 2 Presents a visual overview of the data generation process using GAN variants. As shown in step 1, the data is divided into training and test sets, and then XGBoost and SVM classifiers are used for evaluation. In step 2, data is generated from the training set using various selected GAN variants. Step 3 involves the creation of a mixed dataset through the combination of original and synthetic data. Finally, in steps 4 and 5, both datasets-the synthetic and the mixed datasets-are subjected to evaluation using classifiers

bility, and portability, making it suitable for application to tabular and structured data. It can be used for classification and regression tasks.

2.2.2 Support Vector Machines Classifier

Support Vector Machine (SVM) [44] is a commonly used supervised machine learning technique for efficiently handling imbalanced data. SVM is commonly applied to solve both classification and regression problems. Notably, it demonstrated that the SVM classifier is unaffected by dataset class imbalance compared to other algorithms. The core concept of the SVM approach is to find an optimal separation or 'hyperplane' using the kernel to separate two or more various classes and create a rigid boundary between the samples, which will assist in classification and regression.

3 Experimental Methodology

The workflow of this study, as shown in Figure 2, involves a series of well-defined phases. The first step is to divide the original datasets into two subsets: the training set and the test set. The training set gets 80% of the data, and the test set gets 20%. The train data are used to instruct the generative models, specifically for training the generator and discriminator networks of each GAN architecture to generate synthetic data. After the generation of data by GAN variations, the quality of the synthetic data is evaluated by comparing the results obtained with classification algorithms against those obtained using the original data.

3.1 Hyperparameter selection

This section describes the hyperparameter tuning procedure for the selected GAN models. Since GANs are difficult to train and can be sensitive to hyperparameters [45]. Several batch sizes, noise dimensions, numbers of epochs, and learning rates are tested. Regarding the validation schema, we have used stratified K-fold cross-validation.

- **Batch Size** We conducted experiments with various values, including 16, 32, 64, 100, 128, and 264 to determine the optimal batch size.
- **Noise Dimension** We evaluated the effect of varying the noise dimension values of 50, 64, 100, 256, 264, 300, and 350 on training and testing across numerous GAN variants and datasets. Based on training and testing performance, we determined the optimal noise dimension size for each GAN variant and dataset.
- **Number of Epochs** During training, the number of epochs indicates the number of times the entire dataset is presented to the GAN model. We experimented with epoch numbers from 100 to 1000 to ensure satisfactory results without overfitting. More information about the number of epochs will be discussed in section 5: Results and Discussion.
- **Learning Rate** The learning rate is a crucial hyperparameter that affects the weight adjustment step size of the GAN models during training. We tested multiple learning rates to fine-tune the models, including 0.01, 0.001, 1e-4, 1e-5, and 5e-6. These values were chosen based on their effect on training stability, convergence speed, and

Table 1 Hyperparameters Used for GAN Variants

Parameters	Range of Values
Batch Size	16, 32, 64, 100, 128, 264
Noise Dimension	50, 64, 100, 256, 264, 300, 350
Number of Epochs	Range from 100 to 1000
Learning Rate	0.01, 0.001, 1e-4, 1e-5, 5e-6
Stratified K-fold Cross-Validation	5-folds, 10-folds

overall performance. In general, 1e-5 and 5e-6 worked well on our datasets. After conducting a series of experiments across a range of learning rate values, we found that the optimal learning rate value is 5e-6 for the BCW and LC datasets for selected GAN variants. For the CTG dataset, both the learning rate values of 1e-5 and 5e-6 performed well according to different GAN variants.

Table 1 provides a summary of the experimental hyperparameter values. The hyperparameters include batch size, noise dimension, number of epochs, learning rate, and stratified K-fold cross-validation. The values corresponding to each hyperparameters illustrate the range and options investigated during the experiments. Tables A1 and A2 in the Supplementary Material section show a detailed summary of optimal hyperparameters for each GAN model, classifier, and dataset.

3.2 Evaluation Metrics

Based on the confusion matrix, we evaluated the performance of our model using various metrics, including accuracy, sensitivity, specificity, which measures the ability of the model to accurately identify negative instances, and F1-score, which is the harmonic mean of precision and recall. These measures are shown in Equations 1-5.

$$\text{Accuracy} = \frac{(\text{TP}) + (\text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{1}$$

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \tag{2}$$

$$\text{Sensitivity (Recall)} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \tag{3}$$

$$\text{Specificity} = \frac{(\text{TN})}{(\text{TN} + \text{FP})} \tag{4}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

3.3 Dataset

The dataset comprises various types of digital data, including numerical, categorical, time-series, and text data. The quantity of data significantly impacts the quality of implementing effective algorithms for machine learning models. For this paper, we used the Breast Cancer Wisconsin (Diagnostic) and the Fetal Cardiotocography(CTG) datasets, which are publicly available on the UCI Machine Learning Repository, and the Lung Cancer Patient dataset is available on Kaggle. Table 2 provides a detailed summary of the original datasets before removing any features, offering essential information such as the total number of cases, features, and the distribution of the classes within the datasets. First, when preparing the data, we checked if there were missing or duplicated values to handle. Moreover, eliminating irrelevant or redundant features can improve the performance of the model [46] and reducing the dimensionality of a dataset by dropping less informative features can improve computational efficiency and reduce the risk of overfitting [47]. To understand the correlation relationship between features, we refer to Section 4.

- **Breast Cancer Wisconsin (BCW)** Breast Cancer Wisconsin (Diagnostic) consists of 569 patients with breast tumours, of which 212 cases are malignant, and the remaining 357 cases are benign. Thirty-two features characterise the tumours. Three properties represent each feature: mean, standard error, and worst value and the features are specified by real values, except for the label, which is categorical. We noticed that the BCW dataset had no missing values, but the last column was empty, and the Id column was redundant and not useful, so we had to drop them. Also, we dropped features that are highly correlated with each other; redundant features can be candidates for removal. More details related to the dropping features are provided in Supplementary Material, and Table A3 in the Supplementary Material section shows a list of features that were dropped.
- **Lung Cancer Patient (LC)** The Lung Cancer dataset contains 1000 records and 25 features indicating the symptoms and risk levels (low, medium, and high) associated with factors related to lung cancer. We used 24 features and dropped the Patient Id column. The features are scaled on either a (1-7), (1-8), or (1-9) scale, where 1 represents the minimum level and 7, 8, and 9 represent the maximum level. Table A3 in the Supplementary Material section shows the list of dropped features.
- **Fetal Cardiotocography(CTG)** The Cardiotocography (CTG) dataset consists of Fetal Heart Rate (FHR) and Uterine Contraction (UC) data classified by medical professionals. It encompasses 2,126 fetal cardiotocogram samples that were autonomously processed. These samples are categorised into 1655 normal, 295 suspicious, and 176 pathologic samples. Researchers can use this dataset to explore both 10-class and 3-class classification problems. The original raw data of the CTG dataset

Table 2 Datasets Overview

Name of Dataset	Total Available Cases	Attributes	Distribution of Classes
Breast Cancer Wisconsin	569 Patients	32	Two imbalanced classes (Malignant(M)=357, Benign(B)=212)
Lung Cancer Patient	1000 Records	25	Three classes (Low, Medium, and High-risk levels); Low-risk level=303, Medium risk level class=332, High-risk level class=365
Cardiotocography (CTG)	2126 Samples	40	Three imbalanced classes; Normal Cases Class(N)=1655, Suspicious Cases Class(S)=295, Pathological Cases Class(P)=176

consists of 40 features. Two available versions of the dataset exist publicly, with one containing 21 features and another containing 23. Different researchers have used different numbers of features; some used 21 [48–50] and others 23 [51–53] although not all features are deemed equally important, ten features are essential features [54]. In this study, we opted for 23 features. Table A3 in the Supplementary Material section provides the list of dropped features. Table A4 shows a coparrission between the sizes of each dataset accross different GAN models.

4 Results and Discussion

This section presents a comparative study and discusses the experiments conducted primarily on different datasets. The experiments are designed to investigate the general properties and performance of various GAN models for generating synthetic data in the clinical domain. The environmental setup for the experiments includes Python 3.7.12, Colab and Kaggle Notebook. We employed the YData-Synthetic package [55] to implement GAN, CGAN, CRAMER GAN, DRAGAN, and WGAN. For CTGAN implementation, we utilised the pre-existing CTGAN [56] library. Additionally, we used other standard Python libraries such as Pandas, NumPy, Random, Seaborn, Matplotlib, and Scikit-learn.

We calculated correlations between real and synthetic datasets to understand variable relationships and identify outliers. For the BCW dataset, we utilised Pearson correlation [57] due to its continuous nature. On the other hand, for the LC and CTG datasets, we applied Spearman's rank correlation [58] due to their ordinal values. This tailored approach allowed for a nuanced analysis, considering the distinct characteristics of each dataset.

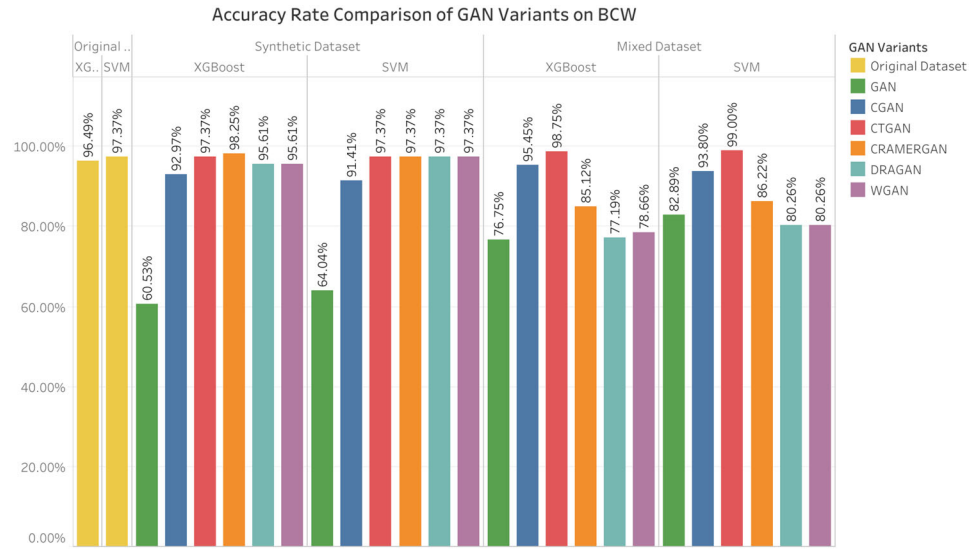
Finally, to implement the statistical test, we used the Statistical Tests for Algorithms Comparison (STAC) [59] platform, which performs statistical analysis to compare outcomes produced by computational intelligence algorithms. It is publicly accessible from the STAC webpage. The implementations of the GAN models presented in this paper are freely accessible in the GitHub repository (https://github.com/Halal-Abdulrahman-Ahmed/MedSynth_GANVariants).

4.1 Breast Cancer Wisconsin Dataset

The experimental results gained from the GAN models on the BCW dataset exhibited outstanding performance on synthetic datasets, with remarkably similar outcomes. Recent studies have highlighted the potential of GAN models to generate data that closely resembles the original data, but this does not always lead to improved classifier performance [60].

Through the conducted experiments, we observed in Figure 3 that CRAMER GAN, DRAGAN, and WGAN outcomes significantly dropped when implemented on mixed datasets. The accuracy of the classifiers using mixed datasets decreased as more training data was provided, indicating potential overfitting. This may be because GAN models are trained to generate data similar to the training data, not data representative of the real world, this means that GAN-generated data can sometimes be misleading. Classifier performance on mixed datasets is lower than on synthetic datasets because GANs are more susceptible to overfit to particular patterns in the original data. This is due to the fact that the presence of synthetic data can impede the ability of GAN models to distinguish between real and fake data, which can lead the model to struggle to generalise and learn patterns specific to the training data, leading to a drop in performance.

Interestingly, CGAN and CTGAN demonstrated exceptional performance, even on the mixed dataset, thereby highlighting their effectiveness in generating synthetic data, as illustrated in Figure 3. Moreover, CTGAN was more suitable than other GAN models for generating BCW data



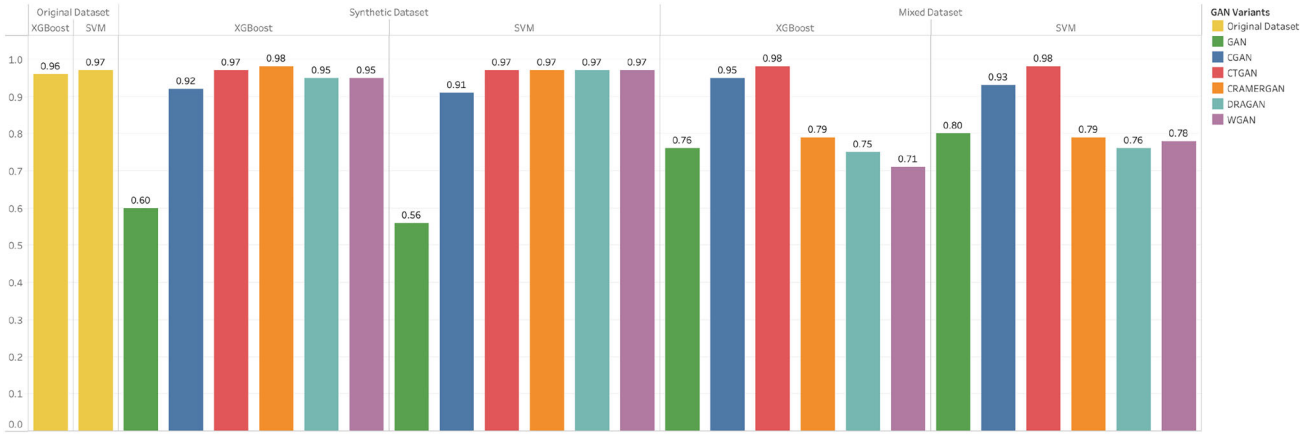


Fig. 3 Comparative Accuracy of GAN Variants with XGBoost and SVM classifiers on BCW Datasets

Fig. 4 Comparative Analysis of F1-scores for GAN Variants on BCW Datasets

because it is designed to generate tabular data due to its conditional generation approach, unlike other GAN variants with more general-purpose applications. The F1-score comparison in Figure 4 provides a comprehensive overview of the experiment outcomes, offering insights into the precision and recall trade-off. F1-score, which is a balanced measure of a model's overall performance, is critical in machine learning evaluations.

Figure 5 shows the Pearson correlation calculated on both the original BCW dataset and the data generated by CTGAN, providing valuable insights into the relationships between features. The heatmap for the original dataset demonstrates numerous strong positive correlations among features. Notably, there are no strong negative correlations, suggesting a lack of consistent inverse relationships between features. If we take the diagnosis variable as an example, we observe strong correlations, particularly with parameters like `parameter_ mean`, `compactness_mean`, `concavity_mean`, and `concave_point_worst`. We could say that diagnosis is the most critical variable, as this variable shows whether the person has breast cancer or not, so we could say that the best GAN model is the model that can capture the same correlation between diagnosis and those features. As we can observe, the diagnosis feature shows the same pattern, i.e. we can observe in the synthetic dataset generated by CTGAN these strong correlations mentioned before between diagnosis and `parameter_mean`, `compactness_mean`, `concavity_mean`, and `concave_point_worst`.

In the CTGAN-generated data, the diagnosis variable demonstrates strong positive correlations with the following features: `concavity_mean`, `area_se`, `compactness_mean`, and `concave_point_worst`. Three out of the top four correlations between the diagnosis variable and these features are effectively captured by the CTGAN model. The correlation coefficients for CTGAN are 0.74, 0.77, and 0.78, whereas the corresponding values for the original dataset are 0.60, 0.70, and 0.79. Additionally, CTGAN replicates the

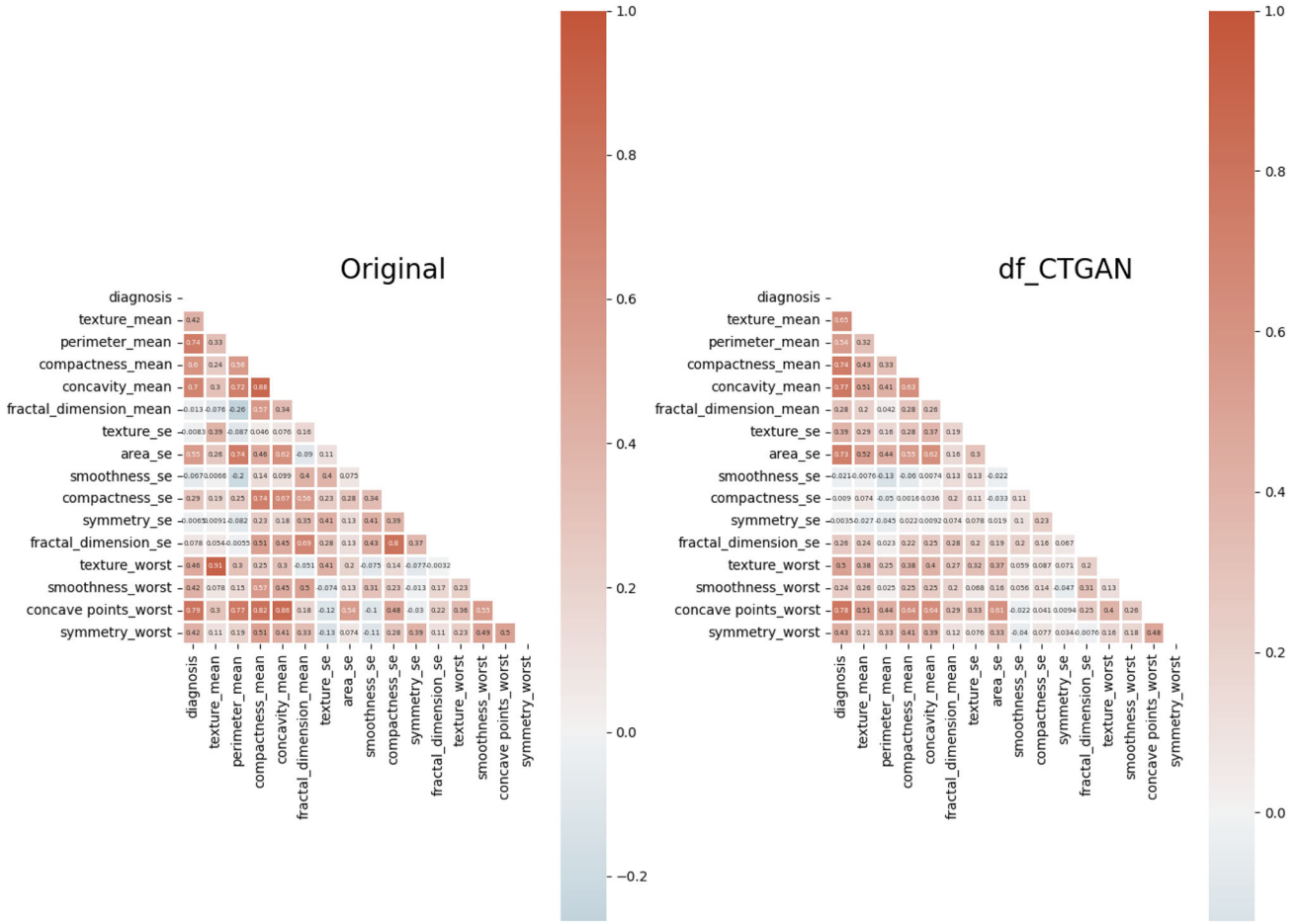


Fig. 5 Heatmaps for BCW Datasets (Original and Synthetic Datasets)

correlation in the `area_se` feature more accurately than the original dataset, with values of 0.73 versus 0.55, respectively.

The correlations in data generated by the CGAN are presented in Figure A1 in the Supplementary Material section. Among the top four features with the strongest correlations to the diagnosis, CGAN appears to replicate all features but with a notable increase in linearity between them. Finally, GAN, DRAGAN, CRAMERGAN, and WGAN could not capture the correlation between diagnosis and other features, as shown in Figure A1 in the Supplementary Material section.

4.2 Lung Cancer Patient Dataset

The experimental results showed that CGAN and CTGAN demonstrated good performance, showcasing good results in the classifiers using synthetic datasets and even outperforming other models on the mixed dataset, see Figure 6. CTGAN emerged as the most successful, exhibiting consistent performance across synthetic and mixed datasets, as shown in Figure 6. On the other hand, GAN, CRAMER GAN, DRAGAN, and WGAN showed relatively weaker performance, mainly when applied solely to synthetic datasets. Surprisingly, there was a frontier improvement in performance on the mixed dataset; however, these models still needed to achieve satisfactory results. These results can be a consequence of hyperparameter tunning issues. Figure 7 represents comparative F1-score outcomes achieved from evaluating different GAN models.

Figure 8 compares the original dataset with the synthetic data from CRAMER GAN and WGAN-generated datasets. It is evident from Figure 8 that the generated synthetic data contains a significant number of outliers. The data generated by CRAMER GAN does not exhibit strong correlations, indicating that changes in one variable are not consistently associated with changes in the other. In Spearman's rank correlation heatmap of WGAN, we observe no correlation between the corresponding pair of features and no monotonic relationship between the ranks of the values in features. Additionally, some columns in the generated data consist of random values. This randomness might result in a lack

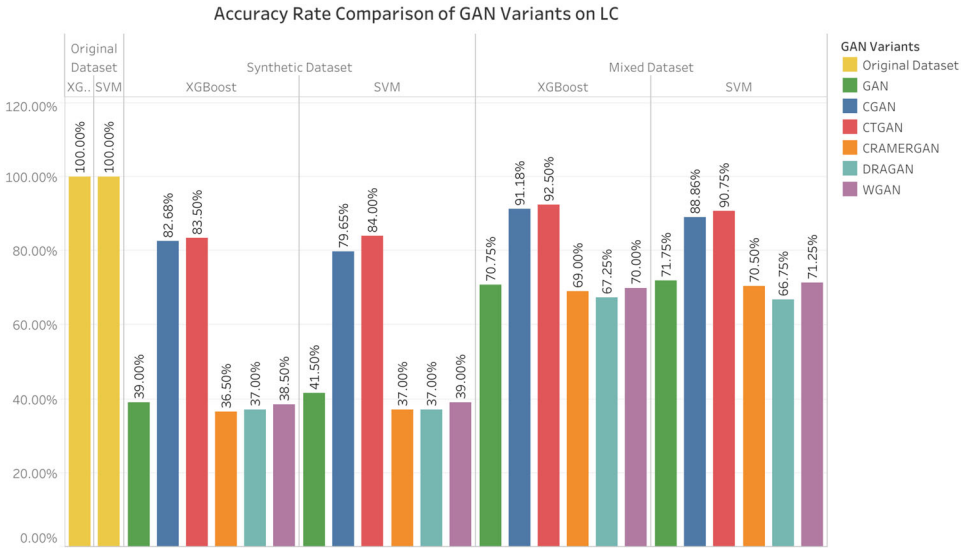


Fig. 6 Comparative Accuracy of GAN Variants on LC Datasets

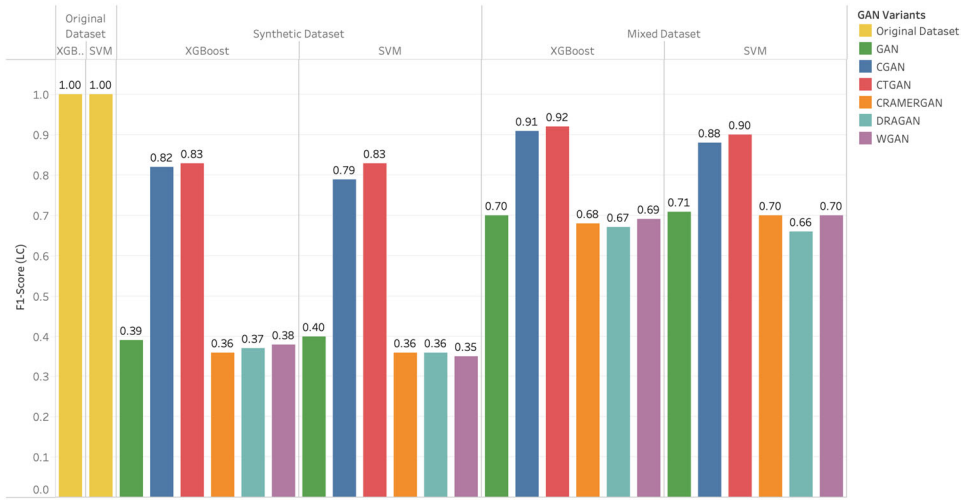


Fig. 7 Comparative Analysis of F1-scores for GAN Variants on LC Datasets

of structured relationships between the variables in those columns.

In this study, some features in the generated synthetic data do not have the same distribution as in the original dataset. This indicates a lack of fidelity in replicating the statistical patterns of the original features. The mentioned issues, such as outliers, mismatched feature distributions, and the absence of correlation between features, negatively impacted the performance of the GAN, CRAMER GAN, DRAGAN and WGAN on the LC dataset. Figure A1 in the Supplementary Material section illustrates a group of Spearman's rank correlation heatmaps that compare original and synthetic datasets generated by selected GAN models.

4.3 Fetal Cardiotocography(CTG) Dataset

Figures 9 and 10 visually represent the experimental results on the CTG dataset. Noteworthy findings indicate satisfactory performance in terms of accuracy and F1-score across the classifier's performance using the data generated by the most GAN models. Remarkably, CGAN demonstrated the most exceptional level of performance. CGAN reveals remarkably improved accuracy, exceeding the original dataset compared to the other selected GAN models. This highlights that introducing additional information, such as class labels, into the CGAN can dramatically increase the accuracy of the classifiers. We notice a slight drop in the performance on mixed datasets, probably due to the overfitting of the GAN models. The CTG dataset is more complex than the other datasets used in this study. The complexity derives from the nature of the features within the CTG dataset, which has many binary and ordinal features, and this can contribute to the risk of overfitting in a model.

However, it is possible to generate synthetic data with satisfactory accuracy results while obtaining less favourable correlation results. In machine learning, particularly in gen-

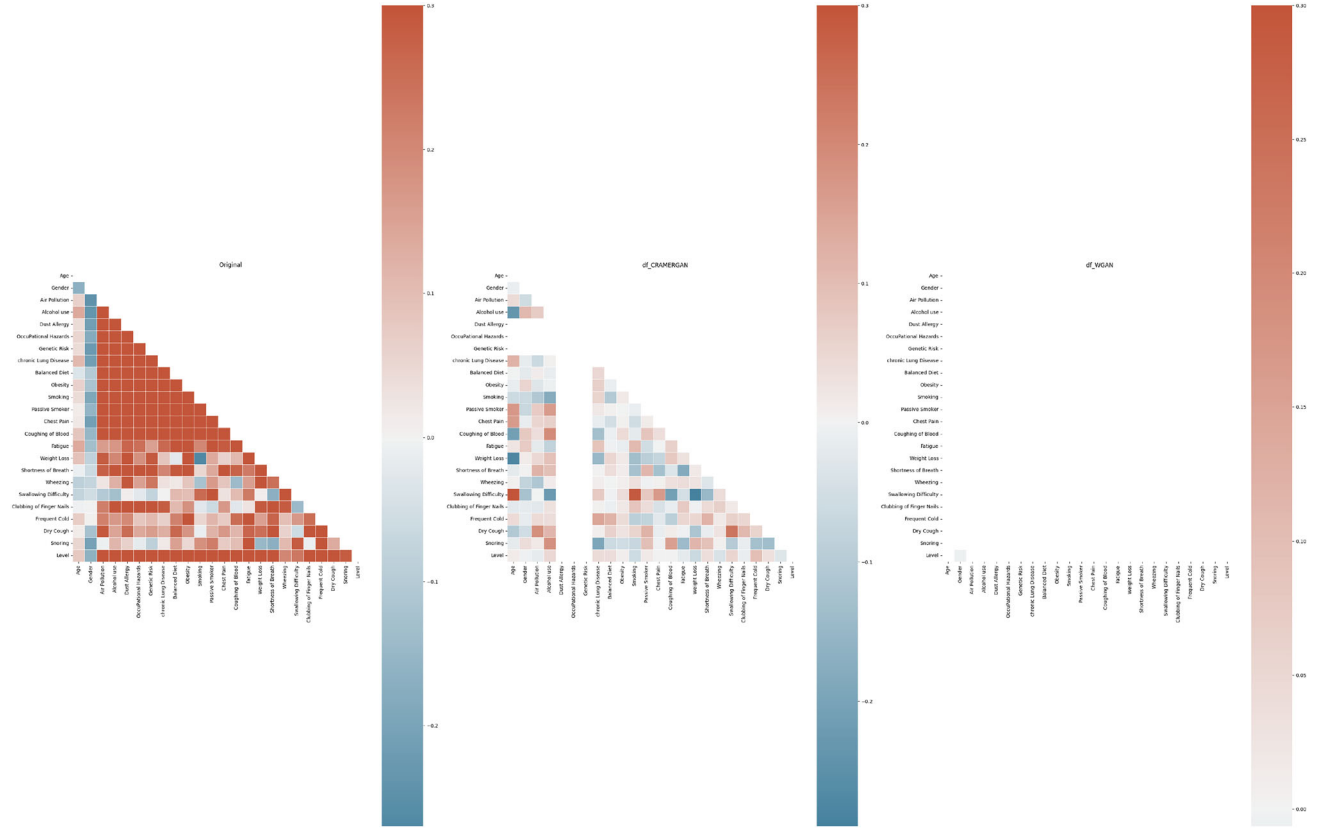


Fig. 8 Spearman's Rank Correlation Heatmaps: Original and Synthetic Datasets (CRAMER GAN and WGAN)

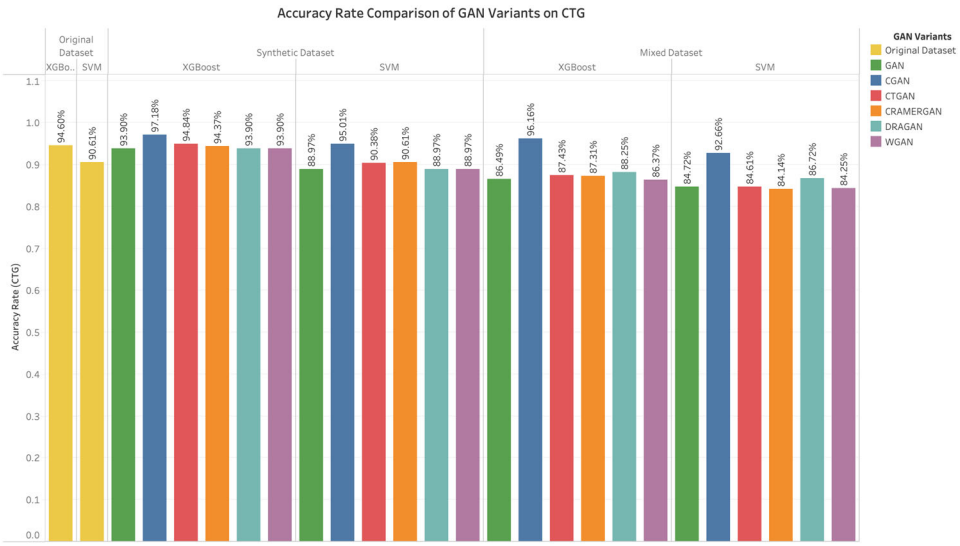


Fig. 9 Comparative Accuracy of GAN Variants on CTG Datasets

erative models, the goal is often to generate synthetic data. Nevertheless, the effectiveness of the model is evaluated based on how well it replicates the underlying patterns and relationships that exist in the original dataset. For this reason, we calculated Spearman's rank correlation. As shown in Figure A3 in the Supplementary Material section, CGAN and CTGAN revealed the ability to preserve a consistent pattern while capturing monotonic relationships between the features of the original data and the generated data. However, it is noteworthy that CTGAN faced challenges in fully replicating every monotonic relationship. In contrast, CGAN succeeded in preserving all relationships. Consequently, based on the evaluation using Spearman's rank correlation, CGAN is considered the more effective model in its capacity to capture

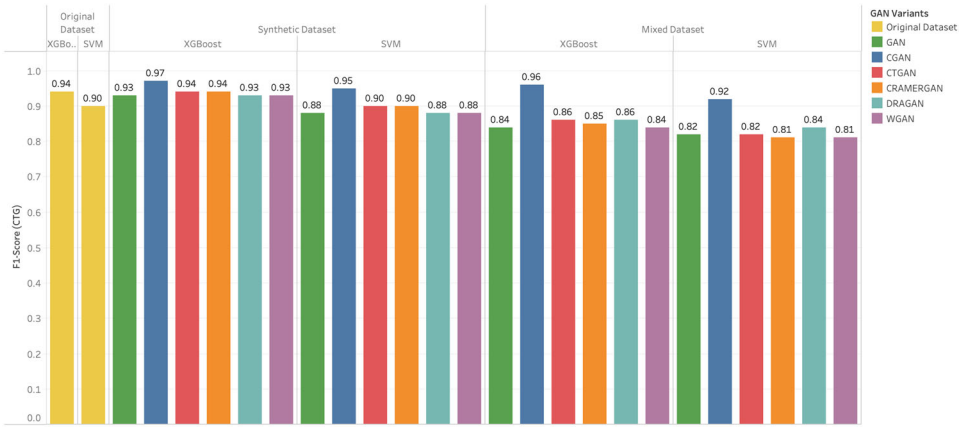


Fig. 10 Comparative Analysis of F1-scores for GAN Variants on CTG Datasets

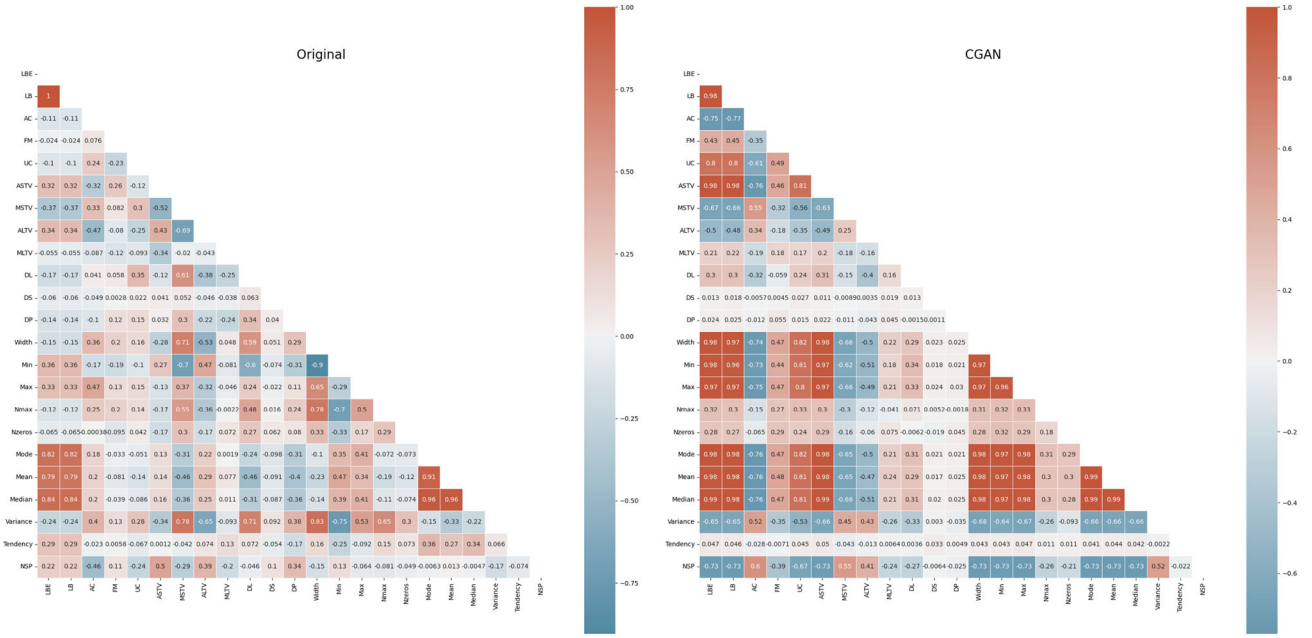


Fig. 11 Spearman's Rank Correlation Heatmaps: Original and Synthetic Dataset (CGAN)

the intended relationships, as shown in Figure 11, which compares the original and generated data with CGAN. From Figure 11, we can observe that the correlation coefficient of features Median and Mode in the original dataset is 0.96, whereas in the CGAN-generated dataset is 0.99, this indicates that the CGAN has successfully generated synthetic data.

4.4 Statistical Test

When dealing with multiple algorithms and classifiers, assessing their effectiveness becomes challenging, especially when performance differences are minor (less than 1%) [61]. A more robust analysis is required for small performance gaps to determine whether one model is genuinely better. Statistical methods, such as hypothesis testing, become essential for establishing the significance of performance differences [62]. When dealing with numerous models, statistical tests like the Friedman test [38] can provide insights into overall performance and identify if one model stands out. Despite the fact that accuracy is a common metric, other factors like precision, recall, or F1-score should also be considered in a comprehensive evaluation. In this study, to conduct the Friedman test, we chose four evaluation metrics as our references: Accuracy, F1 Score, Recall (Sensibility), and Specificity. The Nemenyi test [63] can be employed to identify pairwise groups that differ significantly (using rank sums or rank means) when the Friedman test yields a significant result. For configuring the settings of the STAC [1] platform, careful consideration was given to ensure a robust and reliable analysis of algorithmic performance. For post-hoc comparisons, we opted the Nemenyi test as a suitable method. Additionally,

[1] http://tec.citius.usc.es/stac

Table 3 Top-Ranked GAN Models for Synthetic Dataset Generation

Rank	BCW Dataset	LC Dataset	CTG Dataset
1st	CTGAN	CTGAN	CGAN
2nd	CGAN	CGAN	CTGAN

we chose a significance level often denoted as alpha (α) of 0.05 [64], as the threshold for determining statistical significance. This signifies that the researchers are willing to accept a 5% probability of making a Type I error (rejecting a true null hypothesis) [65].

The Friedman test, performed solely on synthetic data generated by each selected GAN model, was included in the statistical analysis. This ensures the test focuses only on comparing the effectiveness of the synthetic datasets generated by each model without affecting the leverage of real-world data.

As depicted in Table 3, the outcomes of the Friedman test highlight the performance of these top-ranking GAN models across different datasets. Concerning the BCW dataset, CTGAN outperformed GAN models; however, the difference between CGAN and CTGAN was minimal Similar to BCW, CTGAN emerged as the superior performer for LC dataset. Furthermore, when considering the CTG dataset, most of the GAN models performed well, but CGAN demonstrated superior performance. This result underscores the effectiveness of GAN models, which may vary based on the specific characteristics of the dataset.

To visualise the performance of the top-ranking GAN models, we present a Critical Difference (CD) diagrams of average score ranks derived from the Friedman test and the Nemenyi post-hoc test. As observed in the CD diagram, CTGAN and CGAN demonstrate superior performance compared to the other GAN variants. The CD diagram ranks the models, where non-overlapping intervals indicate significant differences in performance. Higher-ranking models, which are considered superior in generating synthetic data are ranked from right to left, as shown in Figures 12 13 and 14.

4.5 Time Usage

The training and testing of each GAN model were meticulously performed across epochs ranging from 100 to 1000 to determine its peak performance epoch. The results, illustrated in Figure 15, summarise the epoch numbers for each GAN model across the datasets. For instance, the CTGAN model demonstrated outstanding performance at 1000 epochs on the synthetic BCW dataset using the XGBoost classifier and 700 epochs with the SVM classifier on the same dataset. Furthermore, at 300 epochs, both XGBoost and SVM achieved optimal performance on the synthetic LC dataset. In addition, the synthetic dataset of CTG demonstrated optimal performance at 100 epochs when using the XGBoost classifier and at 1000 epochs when using the SVM classifier. When analysing the epoch numbers of GAN variants, we have discovered notable variations in performance patterns depending on the specific classifier and dataset employed. The dynamic interaction between GAN models and epoch numbers highlights the significance of tuning training hyperparameters to optimise the quality of synthetic data for various datasets and models. This enables an informed choice of the most effective epoch for every GAN variant.

5 Conclusions

This study evaluated six GAN models-GAN, CGAN, CTGAN, CRAMER GAN, DRAGAN, and WGAN-for generating synthetic tabular medical data. To assess the effectiveness of these models, we employed the Friedman test to identify statistically significant differences among them and used post hoc pairwise comparisons based on rank sums or rank means. In addition, we used Pearson's correlation for continuous data and Spearman's correlation for ordinal data to assess the similarity between real and synthetic datasets for a couple of reasons; Linear correlation analysis is a widely used technique to assess how well synthetic data preserve relationships between features [66–69]. On the other hand, previous studies have also used linear correlation to evaluate synthetic data quality [67, 69, 70]. Based on this, we followed a similar approach to ensure consistency with existing research. However, other similar metrics, such as coverage and structural similarity, exist and were not within the scope of this study, but for future work, additional metrics could be explored. For the classification tasks, we used two different classifiers, XGBoost and SVM. The chosen datasets are Breast Cancer Wisconsin (Diagnostic), Lung Cancer Patient, and Fetal Cardiotocography CTG. The results are also slightly different since the statistical properties of the three given datasets differ regarding the number of samples, features, data type, and distribution.

It is essential to highlight the importance of tailoring GAN architectures to specific datasets. The tabular data generated by more advanced GAN models, such as CGAN and CTGAN, significantly improved accuracy. Our experiments revealed that CTGAN showed promising results in generating tabular data of BCW and LC datasets and CGAN of the CTG dataset. This research opens avenues for future exploration in the generation of synthetic tabular data, including advanced techniques to improve the quality and diversity of generated datasets. Expanding this work, future research may explore the use of a hybrid approach that combines the

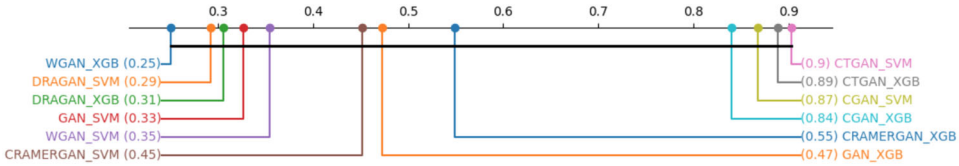


Fig. 12 Critical Difference (CD) Diagram of Average Score Ranks for the Breast Cancer Dataset

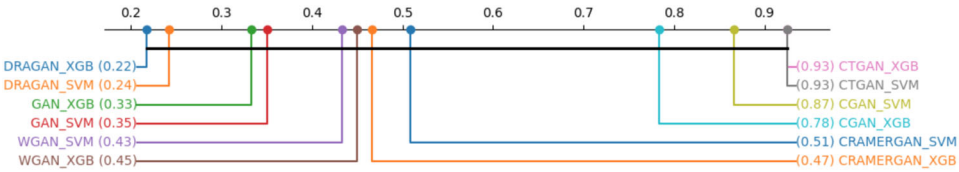


Fig. 13 Critical Difference (CD) Diagram of Average Score Ranks for the Lung Cancer Dataset

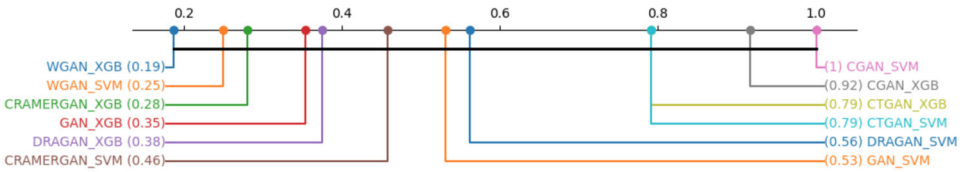


Fig. 14 Critical Difference (CD) Diagram of Average Score Ranks for the CTG Dataset

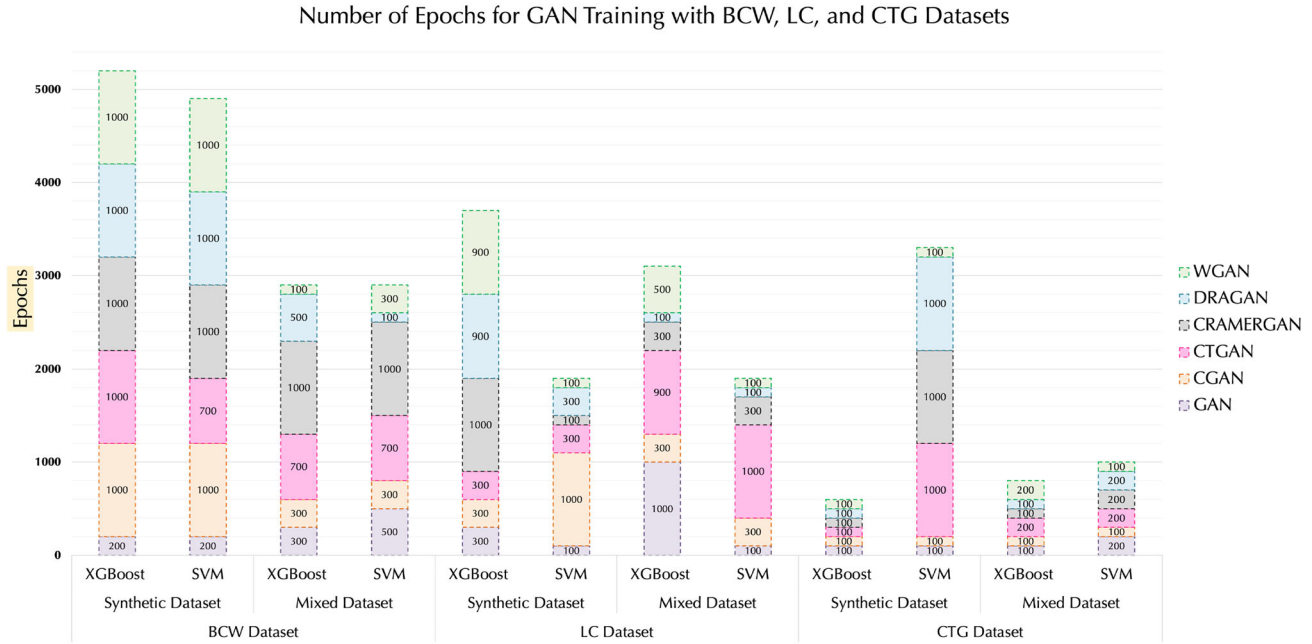


Fig. 15 This graphical representation provides a brief comparison of GAN training across three distinct datasets: BCW, LC, and CTG. It highlights the training epochs required for various GAN variants when evaluated with two classifiers, XGBoost and SVM, across synthetic and mixed datasets. The y-axis displays the number of training epochs, with specific counts for each GAN variant

strengths and overcomes the deficiency of two independent models that include resampling techniques and synthetic data generation methods

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41060-025-00816-w.

Acknowledgements This work has been funded by the Spanish Ministry of Science and Innovation under project PID2020-117954RB-C22 financed by MCIN/ AEI /10.13039/501100011033.

Author Contributions HAA carried out the analysis and prepared the manuscript. JN and INCH provided the paper concept, conceived the methods, and led the project. BVM and HAA performed the experiments. JN, BVM, and INCH reviewed and edited the final manuscript. Finally, all authors certify that they have participated sufficiently in the work to take public responsibility for the content.

Funding Funding for open access publishing: Universidad de Sevilla/CBUA

Data Availability All the code is available on https://github.com/Halal-Abdulrahman-Ahmed/MedSynth_GANVariants. All experiments are conducted using publicly accessible datasets that are included in the previous repository.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethical and informed consent for data used Breast Cancer Wisconsin [31], Lung Cancer Patient [32] and Fetal Cardiotocography [33] datasets are freely accessible from UCI Machine Learning Repository.

References

1. Vallevik, V.B., Babic, A., Marshall, S.E., Severin, E., Brøgger, H.M., Alagaratnam, S., Edwin, B., Veeraragavan, N.R., Befring, A.K., Nygård, J.F.: Can i trust my fake data–a comprehensive quality assessment framework for synthetic tabular data in healthcare. International Journal of Medical Informatics, 105413 (2024)
2. Nik, A.H.Z., Riegler, M.A., Halvorsen, P., Storås, A.M.: Generation of synthetic tabular healthcare data using generative adversarial networks. In: International Conference on Multimedia Modeling, pp. 434–446 (2023). Springer
3. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
4. McDuff, D., Curran, T., Kadambi, A.: Synthetic data in healthcare. arXiv preprint arXiv:2304.03243 (2023)
5. D'amico, S., Dall'Olio, D., Sala, C., Dall'Olio, L., Sauta, E., Zampini, M., Asti, G., Lanino, L., Maggioni, G., Campagna, A., et al.: Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clinical Cancer Informatics **7**, 2300021 (2023)
6. Vacca, J.R.: Computer and Information Security Handbook, 2nd edn. Newnes, Burlington, MA (2012)
7. Pieters, M., Wiering, M.: Comparing generative adversarial network techniques for image creation and modification. arXiv preprint arXiv:1803.09093 (2018)
8. Torres-Reyes, N., Latifi, S.: Audio enhancement and synthesis using generative adversarial networks: A survey. International Journal of Computer Applications **182**(35), 27–31
9. Murtaza, H., Ahmed, M., Khan, N.F., Murtaza, G., Zafar, S., Bano, A.: Synthetic data generation: State of the art in health care domain. Computer Science Review **48**, 100546 (2023)
10. Han, C., Murao, K., Noguchi, T., Kawata, Y., Uchiyama, F., Rundo, L., Nakayama, H., Satoh, S.: Learning more with less: Conditional pggan-based data augmentation for brain metastases detection using highly-rough annotation on mr images. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 119–127 (2019)
11. Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J.: Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, pp. 732–740 (2018). Springer
12. Bhagat, V., Bhaumik, S.: Data augmentation using generative adversarial networks for pneumonia classification in chest xrays. In: 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 574–579 (2019). IEEE
13. Uzunova, H., Ehrhardt, J., Jacob, F., Frydrychowicz, A., Handels, H.: Multi-scale gans for memory-efficient generation of high resolution medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, pp. 112–120 (2019). Springer
14. Munia, M.S., Nourani, M., Houari, S.: Biosignal oversampling using wasserstein generative adversarial network. In: 2020 IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–7 (2020). IEEE
15. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)
16. Li, J., Cairns, B.J., Li, J., Zhu, T.: Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. NPJ Digital Medicine **6**(1), 98 (2023)
17. Mottini, A., Lheritier, A., Acuna-Agost, R.: Airline passenger name record generation using generative adversarial networks. arXiv preprint arXiv:1807.06657 (2018)
18. Azman, M.S., Rossi, F., Zulkarnain, N., Mokri, S.S., Abd Rahni, A.A., Ali, N.F.: Classification of lung nodule ct images using gan variants and cnn. In: 2022 IEEE International Conference on Computing (ICOCO), pp. 310–315 (2022). IEEE
19. Chin-Cheong, K., Sutter, T., Vogt, J.E.: Generation of heterogeneous synthetic electronic health records using gans. In: Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (2019). ETH Zurich, Institute for Machine Learning
20. Hussain, B.Z., Andleeb, I., Ansari, M.S., Joshi, A.M., Kanwal, N.: Wasserstein gan based chest x-ray dataset augmentation for deep learning models: Covid-19 detection use-case. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2058–2061 (2022). IEEE
21. Baowaly, M.K., Lin, C.-C., Liu, C.-L., Chen, K.-T.: Synthesizing electronic health records using improved generative adversarial networks. Journal of the American Medical Informatics Association **26**(3), 228–241 (2019)
22. Rashidian, S., Wang, F., Moffitt, R., Garcia, V., Dutt, A., Chang, W., Pandya, V., Hajagos, J., Saltz, M., Saltz, J.: Smooth-gan: towards sharp and smooth synthetic ehr data generation. In: Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, Auguest 25–28, 2020, Proceedings 18, pp. 37–48 (2020). Springer
23. Yoon, J., Drumright, L.N., Van Der Schaar, M.: Anonymization through data synthesis using generative adversarial networks (ads-gan). IEEE journal of biomedical and health informatics **24**(8), 2378–2388 (2020)
24. Nasimov, R., Nasimova, N., Mirzakhalilov, S., Tokdemir, G., Rizwan, M., Abdusalomov, A., Cho, Y.-I.: Gan-based novel approach for generating synthetic medical tabular data. Bioengineering **11**(12), 1288 (2024)

25. Alqulaity, M., Yang, P.: Enhanced conditional gan for high-quality synthetic tabular data generation in mobile-based cardiovascular healthcare. Sensors **24**(23), 7673 (2024)
26. Kang, H.Y.J., Batbaatar, E., Choi, D.-W., Choi, K.S., Ko, M., Ryu, K.S.: Synthetic tabular data based on generative adversarial networks in health care: generation and validation using the divide-and-conquer strategy. JMIR Medical Informatics **11**, 47859 (2023)
27. Yadav, P., Gaur, M., Madhukar, R.K., Verma, G., Kumar, P.: Rigorous experimental analysis of tabular data generated using tvae and ctgan. International Journal of Advanced Computer Science & Applications **15**(4) (2024)
28. Fonseca, J., Bacao, F.: Tabular and latent space synthetic data generation: a literature review. Journal of Big Data **10**(1), 115 (2023)
29. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., Rankin, D.: Synthetic data generation for tabular health records: A systematic review. Neurocomputing **493**, 28–45 (2022)
30. Liu, T., Qian, Z., Berrevoets, J., Schaar, M.: Goggle: Generative modelling for tabular data by learning relational structure. In: The Eleventh International Conference on Learning Representations (2023)
31. Wolberg, William, Mangasarian, Olvi, Street, Nick, Street, W.: Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B (1995)
32. DAMARLA, R.: Cancer Patients Data. https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data. [Accessed 2023-01-01] (2020)
33. Campos, D., Bernardes, J.: Cardiotocography. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51S4N (2010)
34. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
35. Gillies, D.B.: Solutions to general non-zero-sum games. Contributions to the Theory of Games **4**(40), 47–85 (1959)
36. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
37. Alqahtani, H., Kavakli-Thorne, M., Kumar, G.: Applications of generative adversarial networks (gans): An updated review. Archives of Computational Methods in Engineering **28**, 525–552 (2021)
38. Bellemare, M.G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., Munos, R.: The cramer distance as a solution to biased wasserstein gradients. arXiv preprint arXiv:1705.10743 (2017)
39. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017)
40. Hong, Y., Hwang, U., Yoo, J., Yoon, S.: How generative adversarial networks and their variants work: An overview. ACM Computing Surveys (CSUR) **52**(1), 1–43 (2019)
41. ArjomandBigdeli, A., Amirmazlaghani, M., Khalooei, M.: Defense against adversarial attacks using dragan. In: 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1–5 (2020). IEEE
42. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp. 214–223 (2017). PMLR
43. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
44. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)
45. Matchev, K.T., Roman, A., Shyamsundar, P.: Uncertainties associated with gan-generated datasets in high energy physics. SciPost Physics **12**(3), 104 (2022)
46. Radha, R., Muralidhara, S.: Removal of redundant and irrelevant data from training datasets using speedy feature selection method. International Journal of Computer Science and Mobile Computing **5**(7), 359–364 (2016)
47. Sachdeva, S., Shi, X.: Dimension reduction. In: Computer Vision, A Reference Guide (2019). https://api.semanticscholar.org/CorpusID:7570591
48. Chamidah, N., Wasito, I.: Fetal state classification from cardiotocography based on feature extraction using hybrid k-means and support vector machine. In: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 37–41 (2015). IEEE
49. Rosario, P.D.: UCI: "Cardiotocography Data Set" - Fetal state's classification – Part 1: Data Summary and EDA — phuongdelrosario.medium.com. https://phuongdelrosario.medium.com/uci-cardiotocography-data-set-fetal-states-classification-part-1-data-summary-and-eda-e0cec8a61eff. [Accessed 26-03-2024]
50. l'Aulnoit, A.H., Parent, A., Boudet, S., Rogoz, B., Demailly, R., Beuscart, R., l'Aulnoit, D.H.: Development of a comprehensive database for research on foetal acidosis. European Journal of Obstetrics & Gynecology and Reproductive Biology 274, 40–47 (2022)
51. Silwattananusarn, T., Kanarkard, W., Tuamsuk, K.: Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble. arXiv preprint arXiv:2010.14051 (2020)
52. Nandipati, S.C.R., XinYing, C.: Classification and feature selection approaches for cardiotocography by machine learning techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC) **12**(1), 7–14 (2020)
53. Ramla, M., Sangeetha, S., Nickolas, S.: Fetal health state monitoring using decision tree classifier from cardiotocography measurements. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1799–1803 (2018). IEEE
54. Bhowmik, P., Bhowmik, P.C., Ali, U., Sohrawordi, M.: Cardiotocography data analysis to predict fetal health risks with tree-based ensemble learning. Inf. Technol. Comput. Sci **5**, 30–40 (2021)
55. GitHub - ydataai/ydata-synthetic: Synthetic data generators for tabular and time-series data — github.com. https://github.com/ydataai/ydata-synthetic. [Accessed 21-01-2024]
56. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. In: Advances in Neural Information Processing Systems (2019)
57. Sedgwick, P.: Pearson's correlation coefficient. Bmj **345** (2012)
58. Sedgwick, P.: Spearman's rank correlation coefficient. Bmj **349** (2014)
59. Rodríguez-Fdez, I., Canosa, A., Mucientes, M., Bugarín, A.: Stac: a web platform for the comparison of algorithms using statistical tests. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2015). IEEE
60. Dat, P.T., Dutt, A., Pellerin, D., Quénot, G.: Classifier training from a generative model. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2019). IEEE
61. Wallis, D.: Comparing classifiers (Friedman and Nemenyi tests) — medium.com. https://medium.com/mlearning-ai/comparing-classifiers-friedman-and-nemenyi-tests-32294103ee12. [Accessed 11-01-2024]
62. Richardson, A.: Nonparametric statistics for non-statisticians: A step-by-step approach by Gregory W. Corder, dale I. foreman. Wiley Online Library (2010)
63. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research **7**, 1–30 (2006)
64. Vega-Márquez, B., Nepomuceno-Chamorro, I.A., Rubio-Escudero, C., Riquelme, J.C.: Ocean: Ordinal classification

with an ensemble approach. Information Sciences **580**, 221–242 (2021). https://doi.org/10.1016/j.ins.2021.08.081
65. Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J., Chaudhury, S.: Hypothesis testing, type i and type ii errors. Industrial psychiatry journal **18**(2), 127 (2009)
66. Bhanot, K., Pedersen, J., Guyon, I., Bennett, K.P.: Investigating synthetic medical time-series resemblance. Neurocomputing **494**, 368–378 (2022)
67. Gonçalves, A., Matos, S., al.: Generation and evaluation of synthetic patient data. Journal of Biomedical Informatics 112, 103611 (2020) https://doi.org/10.1016/j.jbi.2020.103611
68. Soranzo, N., Bianconi, G., Altafini, C.: Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. Bioinformatics **23**(13), 1640–1647 (2007)
69. Wolf, M., Tritscher, J., Landes, D., Hotho, A., Schlör, D.: Benchmarking of synthetic network data: Reviewing challenges and approaches. Computers & Security, 103993 (2024)
70. Vega-Márquez, B., Rubio-Escudero, C., Nepomuceno-Chamorro, I.: Generation of synthetic data with conditional generative adversarial networks. Logic Journal of the IGPL **30**(2), 252–262 (2022)
71. Breast Cancer Machine Learning Prediction — gtraskas.github.io. https://gtraskas.github.io/post/breast_cancer/. [Accessed 18-03-2024]
72. Tumor Diagnosis (Exploratory Data Analysis) — kaggle.com. https://www.kaggle.com/code/harikrishna9/tumor-diagnosis-exploratory-data-analysis#Exploratory-Data-Analysis. [Accessed 18-03-2024]
73. A Study of a Breast Cancer Dataset — Breast Cancer Data Study — ucb-stat-159-s22.github.io. https://ucb-stat-159-s22.github.io/hw07-Group26/README.html. [Accessed 18-03-2024]
74. Sumbria, S.: Breast Cancer Diagnostic Dataset - EDA — medium.com. https://medium.com/analytics-vidhya/breast-cancer-diagnostic-dataset-eda-fa0de80f15bd. [Accessed 18-03-2024]