JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Yusof, Y., & Fajila, F. (2026). Conditional tabular generative adversarial network-based synthetic data generation for model generalisation improvement. *Journal of Information and Communication Technology*, *25*(1), 1-16. https://doi.org/10.32890/jict2026.25.1.1

Conditional Tabular Generative Adversarial Network-based Synthetic Data Generation for Model Generalisation Improvement

[1]Yuhanis Yusof & [2]Fathima Fajila
[1&2]School of Computing, Universiti Utara Malaysia, Malaysia
[2]Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka

[*1]yuhanis@uum.edu.my
[2]fajila@seu.ac.lk
*Corresponding author

Received: 5/5/2025 Revised: 26/6/2025 Accepted: 5/7/2025 Published: 31/1/2026

ABSTRACT

Accessing extensive and varied datasets is essential for developing strong predictive models in data analytics. However, many real-world applications suffer from small and imbalanced datasets, leading to overfitting, poor generalisation, and low model performance. Traditional data augmentation techniques are often unsuitable for tabular data, as they fail to preserve complex feature relationships. To address this challenge, this study adapts the Conditional Tabular Generative Adversarial Network (CTGAN) for synthetic data generation. The proposed approach involves five phases: (1) Data Acquisition, 2) Data Preparation, (3) Model Training, (4) Synthetic Data Generation, and (5) Evaluation. Experimental results on three benchmark datasets show that the proposed work produced data that closely adheres to the statistical distribution of the original dataset, with Wasserstein Distance < 0.05 for numerical features and Jensen-Shannon Divergence < 0.08 for categorical features. Additionally, models trained on datasets including synthetic and real data achieved up to 15% improvement in classification accuracy compared to those trained on real and small datasets alone. Training on a combination of real and synthetic data for the minority class in large datasets significantly improves the F1-score, with gains of approximately 9–10%. This approach also yields a modest increase in overall accuracy (around 1.5%), suggesting enhanced model generalisation. These results indicate that the adapted CTGAN is a viable option for data augmentation, addressing problems with limited and imbalanced data for machine learning data training.

Keywords: Deep learning, CTGAN, data augmentation, synthetic data.

INTRODUCTION

In data analytics, the quality and size of datasets play a crucial role in developing robust and accurate predictive models. Large and diverse datasets enable machine learning models to generalize well to unseen data, reducing overfitting and improving performance. However, data scarcity and imbalance (Sainin et al., 2021) remain significant challenges in many real-world applications. Domains such as healthcare, finance, and cybersecurity often suffer from limited labelled data, making it challenging to train effective models. Traditional approaches, such as oversampling (Wongvorachan et al., 2023; Ahsan et al., 2025) and undersampling (Wongvorachan et al., 2023), are commonly used to address this issue, but may introduce bias or fail to capture complex relationships within the data (Khan et al., 2024). As a result, there is a growing need for advanced data augmentation techniques explicitly tailored for structured tabular datasets.

Unlike image or text data, where augmentation techniques such as flipping, cropping, or token substitution are widely used, tabular data presents unique challenges. Features in structured datasets often have complex dependencies, mixed data types, and non-linear relationships, making traditional augmentation techniques less effective. Popular methods like the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) generate synthetic samples for imbalanced datasets but fail to preserve intricate feature relationships. More recent approaches leverage Generative Adversarial Networks (GANs) (Goodfellow et al., 2020; Gui et al., 2023) to synthesize high-quality data. However, standard GANs struggle with categorical data representation, leading to mode collapse and unrealistic synthetic samples (Chen et al., 2020).

To address these limitations, this study adapts Conditional Tabular Generative Adversarial Network (CTGAN) (Ma et al., 2024; Xu et al., 2019; Zhao & Guan, 2023) as a solution for synthetic data generation in tabular datasets. CTGAN extends the traditional GAN architecture by introducing mode-specific normalization and a conditional vector mechanism, enabling the model to handle both categorical and numerical features effectively. This paper begins by reviewing relevant studies on CTGAN and established data augmentation techniques. Next, it details the procedures for generating and evaluating synthetic datasets. The resulting experimental findings are then examined in depth, before concluding with a summary of the discussion and potential directions for future research.

RELATED WORKS

This section reviews prior research on data augmentation methods, a critical technique for enhancing model performance, particularly with limited datasets. Significant efforts have been dedicated to image augmentation, focusing on geometric and color space transformations to improve deep learning robustness. In natural language processing, text augmentation techniques like back-translation and synonym replacement have been developed to introduce semantic variations. Furthermore, advancements in tabular and audio data augmentation, including SMOTE and CTGAN demonstrate the breadth of ongoing research aimed at maximizing the benefits of data augmentation across diverse data modalities.

Data Augmentation Methods

Data augmentation, a technique employed to increase the size and diversity of training datasets artificially, has become indispensable in modern data analytics, particularly when dealing with small datasets. By introducing variations to existing data, augmentation mitigates overfitting, enhances model robustness, and improves generalisation. This discussion examines data augmentation methods across various data types, including image, text and audio.

For image data, various techniques such as rotation, noise introduction, shearing, and translation have been employed to enhance brain MR images, thereby increasing the image count and improving the effectiveness of several tasks. For instance, a study by Khan et al. (2021) controlled synthesized images' captured environments (e.g., time, lighting) in electric vehicle (EV) charging inlet detection for autonomous EV charging robots. Several studies have utilized random scaling, rotation, and elastic deformation to improve the accuracy of tumor segmentation (Lyu & Tian, 2025; Fidon et al., 2021; Isensee et al., 2021). While these techniques are commonly favored in the literature, augmentation has also been realized by generating synthetic images. The mixup method (Qiu et al., 2021), for instance, creates synthetic images by merging two randomly chosen images along with their corresponding labels.

In textual data, augmentation methods aim to preserve semantic meaning while introducing syntactic variations, thus improving the robustness and generalisation of natural language processing (NLP) models. One widely used method is back-translation, which translates a sentence from the source language into a pivot language (e.g., English ? French ? English) and then back to the original language. This process introduces paraphrases that retain the core message while altering surface-level structure (Sennrich et al., 2016) . To perform this method, neural machine translation (NMT) systems such as MarianMT or Google Translate are employed. However, this method can be computationally expensive, particularly for large datasets or multiple language pairs. Moreover, errors in translation may propagate and result in grammatically incorrect or semantically distorted outputs, especially when using lower-quality translation models or in low-resource languages (Edunov et al., 2018).

Another common category of augmentation involves surface-level lexical alterations such as synonym replacement, random insertion, random deletion, and word swapping. Synonym replacement uses lexical databases like WordNet to substitute a word with its synonym, while the other techniques alter the sentence structure by inserting, deleting, or swapping words randomly. These methods were formalized under the umbrella of Easy Data Augmentation (EDA)(Wei & Zou, 2019), which showed that such approaches significantly boost performance in low-resource classification tasks. Despite their simplicity, these methods suffer from context insensitivity, where the inserted or replaced words may not align well with the surrounding context. For instance, replacing the word "bank" in "river bank" with "financial institution" is inappropriate. Additionally, random deletions or insertions can result in ungrammatical or awkward sentences, potentially introducing noise that harms model performance rather than improving it (Tavor et al., 2020).

More context-aware approaches include contextual word embeddings, such as those derived from models like Bidirectional Encoder Representations from Transformers (BERT) or Robustly Optimized BERT pre-training Approach (RoBERTa). In this method, a word in the sentence is masked, and the model predicts contextually appropriate replacements using its pre-trained language understanding. This technique preserves semantic integrity better than non-contextual methods and is particularly useful for generating plausible and varied augmentations (Kobayashi, 2018). Nevertheless, contextual

augmentation comes with its own set of limitations. The use of large language models requires substantial computational resources, and the augmentation may reflect the biases present in the pre-trained models. Moreover, the substitutions made are often non-transparent and difficult to control, which can hinder interpretability and precision in downstream tasks.

A more recent technique, MixText, augments data in the latent space rather than at the text surface. This method involves interpolating the hidden representations of different input texts to generate synthetic training examples, often used in semi-supervised settings (Chen et al., 2020). While MixText has shown improvements in classification tasks, its implementation complexity is higher than simpler text-based techniques. It also tends to be model-specific, requiring architecture-level integration and thus lacking flexibility for general use across models. Additionally, interpolating between two distinct labels may create examples with ambiguous semantic meaning, posing challenges during supervised learning.

For tabular data, augmentation techniques must be tailored to the nature of numerical and categorical features, each of which poses unique challenges. One of the most common approaches to address class imbalance in classification tasks is the SMOTE. SMOTE generates synthetic examples of the minority class by interpolating between existing samples in feature space, creating new data points along the line segments that connect a sample to its nearest minority class neighbors (Chawla et al., 2002). This technique helps to balance datasets and improve the robustness of classifiers. However, SMOTE may lead to overlapping between classes, particularly in high-dimensional spaces, thereby increasing the risk of overfitting and noise introduction, especially when minority classes are not well-clustered (Fernández et al., 2018). Random permutation of feature values within a column can also be used to break linear dependencies and simulate realistic variation in the data. This technique assumes feature independence and is best suited for non-sequential tabular data. However, indiscriminate shuffling can destroy meaningful correlations between variables (such as age and income), leading to data incoherence (Patki et al., 2016).

Recently, Generative Adversarial Networks (GANs) have gained traction as powerful methods for producing realistic synthetic tabular data, especially for privacy-preserving applications. The CTGAN (Habibi et al., 2023) model addresses the challenges of mixed data types by modeling the distribution of each feature conditionally. CTGAN employs a generator-discriminator framework to synthesize data that closely mirrors the original dataset, making it suitable for both categorical and numerical features (Xu et al., 2019). CTGAN variants have shown success in generating synthetic data that maintains utility for downstream machine learning tasks (Habibi et al., 2023; Ma et al., 2024; Majeed & Hwang, 2023). Despite these advances, GAN-based models are complex to train, requiring careful tuning and often large volumes of data to prevent issues like mode collapse (where the generator produces limited variations).

In audio data, time-domain augmentations alter audio signals' temporal and frequency characteristics, such as time stretching and pitch shifting. Spectrogram augmentations, which manipulate the visual representation of audio, offer another avenue for introducing variations (Iglesias et al., 2023; Uchitomi et al., 2023) in improving speech recognition accuracy. Nevertheless, the selection of augmentation methods depends heavily on the data type and the specific task. Beyond raw audio, spectrogram-based augmentations operate on the visual representation of audio signals, usually a mel spectrogram or log-magnitude spectrogram. A widely adopted technique is SpecAugment, which applies time masking, frequency masking, and time warping to the spectrogram image (Park et al., 2019). These operations simulate signal dropout or occlusion, encouraging models to focus on more general patterns in the audio. However, one key limitation is that these augmentations assume the spectrograms are uniformly

relevant across tasks; for example, frequency masking may inadvertently hide features critical for tone- or pitch-sensitive languages like Mandarin. Moreover, because spectrogram transformations are task-agnostic, they may not capture semantic-level variations required in more nuanced classification tasks such as emotion or intent detection (Dua et al., 2023).

Spectrogram flipping (Zarandah et al., 2023) is an audio augmentation technique that involves horizontally or vertically flipping the spectrogram representation of an audio signal. Horizontal flipping simulates time-reversed audio, which may help models generalize temporal features, while vertical flipping alters the frequency axis, potentially enhancing robustness to pitch variations. This method is easy to implement and can be combined with other spectrogram-based augmentations like masking or warping. However, improper application, especially vertical flipping may distort key acoustic patterns and reduce model performance on tasks sensitive to pitch or frequency content. A recent work on audio augmentation by Tsalera et al. (2025) reports on the impact of dataset augmentation and synthetic generation techniques on the accuracy of supervised audio classification using neural network classifiers. Synthetic sound generation was based on the AudioGen generative model, triggered through a series of customized prompts. Importantly, the selection of augmentation methods must be aligned with the data type and the end-task objectives. For instance, while time stretching may enhance robustness in automatic speech recognition tasks, it may distort temporal cues essential for speaker verification. Therefore, choosing the appropriate augmentation requires domain expertise and empirical validation, especially in audio domains where subtle temporal or spectral cues carry semantic weight.

Deep Learning for Data Augmentation

The increasing reliance on deep learning models for predictive tasks has highlighted the importance of data augmentation as a strategy to overcome the limitations of small or imbalanced datasets. Traditional data augmentation techniques, such as oversampling and noise injection, are often inadequate for complex data structures, leading researchers to explore deep learning-based augmentation methods. By using neural networks to create synthetic data, these techniques maintain statistical features while guaranteeing diversity. Early studies concentrated on text and picture augmentation, where transformations like Variational Autoencoders (VAEs) (Kingma & Welling, 2019) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) showed notable gains in model resilience. However, recent advancements have extended deep learning-based augmentation to structured tabular data, time series, and speech datasets, addressing broader data scarcity and quality challenges.

GANs have been one of the most influential approaches in data augmentation for deep learning models. Introduced by Goodfellow et al. (2020), GANs consist of a Generator-Discriminator pair, where the Generator learns to create synthetic data that mimics real distributions, and the Discriminator differentiates between real and synthetic samples. Variants like Conditional GANs (CGANs) (Hou et al., 2022) and Wasserstein GANs (WGANs) (Biau et al., 2021) have improved the stability of training and enabled class-conditional data generation. The CTGAN (Xu et al., 2019) has been developed specifically for structured data, effectively handling categorical and numerical features in tabular datasets. The success of GAN-based augmentation has been observed across domains such as medical imaging (Zhang et al., 2023), financial modelling (Strelcenia & Prakoonwit, 2023) , and cybersecurity (Dunmore et al., 2023), where high-quality synthetic data enhances model performance.

Beyond GANs, the VAEs have emerged as another powerful deep learning-based augmentation technique. Unlike GANs, VAEs use an encoder-decoder framework to learn a probabilistic latent space from which new synthetic samples can be generated (Kingma & Welling, 2014). VAEs have been widely used in image generation, anomaly detection, and speech synthesis (Sohn et al., 2015; Gabbay & Hoshen, 2019). Various study have explored hybrid VAE-GAN architectures, combining the generative capabilities of VAEs with the adversarial training of GANs to improve synthetic data realism (Larsen et al., 2016). However, a key limitation of VAEs is their tendency to produce blurry samples in high-dimensional spaces, making them less effective for augmentation tasks where fine-grained details matter, such as medical image synthesis.

Conditional Tabular Generative Adversarial Network

The CTGAN (Xu et al., 2019) is an effective tool for generating synthetic tabular data that closely mimics actual datasets. It utilizes the capabilities of GANs (Dunmore et al., 2023) by employing a generator and a discriminator. The generator's goal is to create synthetic data that can deceive the discriminator, while the discriminator's task is to differentiate between real and synthetic data. This competing interaction enhances the generator's skill in producing highly authentic synthetic data. CTGAN integrates several essential features to address the challenges posed by real-world tabular data, including mixed data types, non-Gaussian distributions, and imbalanced categories.

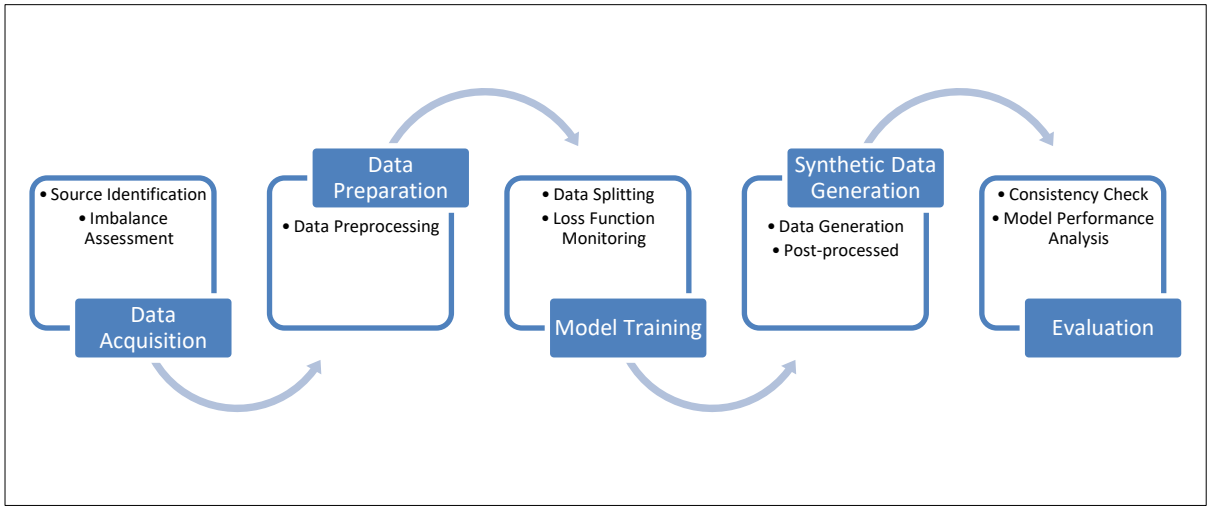
The generator creates synthetic data that closely resembles the statistical characteristics of the original data using random noise as input. Conversely, the discriminator assesses the produced data and establishes its veracity. It provides a score that represents the probability that the data is authentic. CTGAN uses adversarial training, in which the discriminator and generator are taught iteratively, to achieve the learning. To trick the discriminator, the generator learns to create more realistic data, and the discriminator gets better at telling the difference between synthetic and genuine data. CTGAN's ability to generate high-fidelity synthetic data has significant implications for various applications, including privacy preservation, data augmentation, and collaborative data sharing. By enabling the creation of realistic synthetic datasets, CTGAN offers valuable solutions for overcoming challenges associated with sensitive data and data scarcity.

THE PROPOSED METHOD

This study is implemented in five phases: 1) Dataset Acquisition, 2) Data Preparation, 3) Model Training, 4) Synthetic Data Generation, and 5) Evaluation. Figure 1 illustrates the phases along with their main activities.

Figure 1

Methodology Phases and Their Main Activities



Dataset Acquisition

Three popular benchmark datasets were utilized, including the Iris, Heart Disease and Diabetes. Table 1 depicts the properties of the three datasets. The Iris dataset is a small, balanced dataset used for multiclass classification. It contains the measures for three types of Iris: the Setosa, Versicolor, and Virginica. The second dataset (i.e Heart Disease) contains patient health records that determine if a person has heart disease based on risk factors such as age, cholesterol, and blood pressure. The imbalance status for this dataset is considered low, as approximately 54% of the samples are 54% negative and 46% positive. The Diabetes dataset is also known as the Pima Indians Diabetes dataset and it contains health measurements to predict whether a person has diabetes. Its imbalance status is considered moderate to high, as only 35% of the samples are diabetic.

Table 1

Properties of Datasets

Dataset	Samples	Features	Feature Type	Target Classes
Iris	150	4	Numerical	3 (Multiclass)
Heart Disease	303	13	Numerical & Categorical	2 (Binary)
Diabetes	768	8	Numerical	2 (Binary)

Data Preparation

Before applying CTGAN to generate synthetic data, it is essential to preprocess the dataset to ensure compatibility with the model. This includes partitioning the dataset, normalising numerical features, addressing missing values, and encoding categorical variables. Many real-world datasets contain missing values that must be handled before training a model. For example, the Heart Disease dataset has missing values in the Blood Pressure and Cholesterol (mg/dL) features. In this study, the missing values are replaced with mean imputation for numerical features (e.g., cholesterol, blood pressure). And for categorical features (e.g., Chest Pain Type), the study uses mode imputation.

This study also converts categorical features into numerical labels using Label Encoding. For example, in Heart Disease, 'Typical Angina' is encoded as 0, 'Atypical Angina' as 1, 'Non-Anginal Pain' as 2, and 'Asymptomatic' as 3. Numerical data that are normalised to a standard range perform well in the majority of machine learning models. Hence, this study uses Min-Max Scaling to transform numerical features between 0 and 1.

Model Training

The training process begins by splitting the real dataset into 80:20 proportions. The larger portion is used for training while the smaller portion is used for testing. The CTGAN deploys loss function (adversarial learning) by optimizing a binary cross-entropy loss function. CTGAN's training procedure uses the GANs approach, in which the generator and discriminator compete against each other. The generator attempts to minimise loss by creating more realistic samples, whereas the discriminator attempts to maximise loss by properly recognising synthetic data. In this study, training was performed in 100 epochs, where each epoch consists of the following steps:

1. Sample real and synthetic data: A set of real data (X_real) is randomly sampled from the training dataset, while the generator produces a batch of synthetic data (X_synthetic) using random noise as input.
2. Train the discriminator (D): The discriminator is trained to classify samples correctly:
   - It assigns label "1" to real data (X_real) and label "0" to synthetic data (X_synthetic).
   - The loss function is computed based on how well it differentiates real from synthetic.
   - The discriminator is updated using gradient descent to improve its classification ability.
3. Train the generator (G): The generator is updated to create more realistic synthetic data:
   - A new batch of X_synthetic is generated.
   - The loss function is calculated based on whether the Discriminator incorrectly classifies X_synthetic as real.
   - The generator is updated using gradient descent to improve the ability to fool the discriminator.
4. Repeat until convergence: These steps are iteratively repeated across multiple epochs. The training stops when the discriminator's accuracy stabilizes around 50%, meaning it can no longer reliably differentiate real from synthetic data.

Synthetic Data Generation

Once training is complete, the generator creates new synthetic data. To do this:

1. The generator receives random noise samples as input.
2. The generator produces synthetic samples based on the learned data distribution.
3. The generated data is then post-processed:
   - Numerical features are denormalized back to their original range.
   - Categorical features are decoded to their original labels.

Before the final synthetic dataset (refer to Table 2) can be saved and used to augment data in improving machine learning models, it has to undergo a consistency check. The assessment determines whether the synthetic data matched the attributes of the original data. For the Iris dataset, the dataset size has doubled per class, whereas for the Heart Disease and Diabetes datasets, the sample counts per class are approximately equal. On the other hand, the parameter settings of the deployed CTGAN are as in Table 3.

Table 2

Number of Real and Synthetic Data Samples

Dataset	Class	# Real Samples	# Synthetic Samples	Total
Iris	Setosa	50	50	100
	Versicolor	50	50	100
	Virginica	50	50	100
Heart Disease	Negative	164	0	164
	Positive	139	25	164
Diabetes	Negative	500	0	500
	Positive	268	225	493

Table 3

CTGAN Hyperparameter Settings

Hyperparameter	Description	Value
Epochs	Number of times the model sees the full dataset	100
Generator Learning Rate	Step size for updating the weights of the Generator	0.0002
Discriminator Learning Rate	Step size for updating the weights of the Discriminator	0.0002
Discriminator Steps	Number of times the Discriminator updates per Generator update	1

Evaluation

The evaluation was performed solely using the test set as it contains original (real) data, ensuring an unbiased evaluation. There were two training scenarios: 1) train classifiers using only real training data and 2) train classifiers using real + synthetic training data. The accuracy, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were reported only on real test data to measure generalisation. These metrics were produced by five classifiers, which include the Random Forest (RF), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Regression (LR) and Gradient Boosting Machine (GBM). The first three classifiers were utilized due to their strength in 1) determining decision boundaries without overfitting, 2) evaluating if the local geometric structure of the original data has been preserved, and 3) being sensitive to data distribution. On the other hand, the LR was chosen as a baseline method, while the GBM serves as the state-of-the-art method for tabular data. The aim was to investigate whether increasing the dataset size or class-balancing would improve the classification accuracy produced by the deployed classifiers. To further validate the study, two large datasets (having more than 40,000 samples) with high class imbalance were used, and the outcomes were analysed.

The study was conducted on a workstation equipped with an Intel Core i7-10750H (10th Generation) processor, operating on a 64-bit Windows 10 environment. The technical implementation and evaluation were developed using Python 3.9, leveraging TensorFlow 2.10 as the primary deep learning framework for model construction.

RESULTS AND DISCUSSION

Statistical characteristics were examined to assess whether the synthetic data matched the attributes of the original data. This includes using mean, Standard Deviation (SD), Wasserstein Distance and Jensen-Shannon Divergence (JSD).

Table 4

Statistical Comparison of Numerical Features between Real vs. Synthetic Data

Dataset	Feature	Mean (Real)	Mean (Synthetic)	SD (Real)	SD (Synthetic)	Wasserstein Distance
Iris	Sepal Length	5.84	5.81	0.83	0.85	0.042
	Petal Width	1.20	1.19	0.76	0.78	0.031
	Sepal Width	3.06	3.08	0.44	0.45	0.038
	Petal Length	3.76	3.73	1.77	1.79	0.041
Heart Disease	Age	54.37	54.41	9.08	9.12	0.051
	Cholesterol (mg/dL)	246.92	247.25	51.90	52.01	0.047
	Trestbps	131.62	131.75	17.53	17.58	0.130
	Thalach	149.64	149.51	22.90	22.84	0.130
	Oldpeak	1.04	1.06	1.16	1.18	0.020
Diabetes	Glucose Level	121.90	121.57	31.97	32.05	0.035
	BMI	32.45	32.49	6.82	6.85	0.028
	Pregnancies	3.84	3.88	3.37	3.39	0.040
	BP	69.11	69.15	19.35	19.39	0.040
	Skin Thickness	20.53	20.49	15.95	15.98	0.040
	Insulin	79.80	79.84	115.24	115.29	0.040
	Pedigree	0.47	0.49	0.33	0.34	0.020
	Age	33.24	33.29	11.76	11.79	0.050

The data in Table 4 indicate that the synthetic data's means and standard deviations closely match those of the real data. The Wasserstein Distance is less than 0.05, indicating minimal statistical deviation between actual and synthetic distributions. In addition, the Jensen-Shannon Divergence (refer to Table 5) values are also low (< 0.08), confirming that the synthetic dataset preserves categorical feature distributions.

Table 5

Jensen-Shannon Divergence for Categorical Features

Dataset	Categorical Feature	JSD Score (0 = Identical, 1 = Dissimilar)
Iris	Species	0.025
Heart Disease	Chest Pain Type	0.043
Diabetes	Outcome	0.039

Data depicted in Table 6 denotes that by adding CTGAN synthetic data, a higher classification is obtained by all deployed classifiers, where between 3% and 10% increment can be seen across datasets. The highest improvements (up to +10.5%) are observed in the Diabetes dataset, which initially had imbalanced classes. In Heart Disease and Diabetes datasets, synthetic data helps balance class representation, leading to fairer predictions. The F1-score increases significantly, indicating better handling of positive and negative cases.

Table 6

Model Performance on Real vs. Synthetic Data

Dataset	Train Set	Classifier	Accuracy (%)	F1-Score	AUC-ROC
Iris	Real Only	RF	92.10%	0.89	0.91
	Real + Synthetic	RF	96.3% (+4.2%)	0.92	0.95
	Real Only	SVM	89.80%	0.87	0.89
	Real + Synthetic	SVM	93.4% (+3.6%)	0.9	0.94
	Real Only	KNN	87.50%	0.85	0.88
	Real + Synthetic	KNN	91.2% (+3.7%)	0.88	0.92
	Real Only	LR	88.30%	0.86	0.88
	Real + Synthetic	LR	92.5% (+4.2%)	0.89	0.93
	Real Only	GBM	90.40%	0.88	0.9
	Real + Synthetic	GBM	94.8% (+4.4%)	0.91	0.96
Heart Disease	Real Only	RF	78.50%	0.76	0.79
	Real + Synthetic	RF	84.7% (+6.2%)	0.82	0.85
	Real Only	SVM	74.20%	0.72	0.76
	Real + Synthetic	SVM	79.3% (+5.1%)	0.78	0.82
	Real Only	KNN	71.80%	0.69	0.73
	Real + Synthetic	KNN	77.0% (+5.2%)	0.75	0.8
	Real Only	LR	76.10%	0.74	0.77
	Real + Synthetic	LR	81.8% (+5.7%)	0.79	0.83
	Real Only	GBM	79.60%	0.77	0.8
	Real + Synthetic	GBM	85.3% (+5.7%)	0.83	0.87
Diabetes	Real Only	RF	72.00%	0.68	0.74
	Real + Synthetic	RF	82.5% (+10.5%)	0.79	0.84
	Real Only	SVM	68.50%	0.65	0.7
	Real + Synthetic	SVM	78.3% (+9.8%)	0.74	0.81
	Real Only	KNN	65.70%	0.63	0.68
	Real + Synthetic	KNN	75.5% (+9.8%)	0.72	0.78
	Real Only	LR	70.20%	0.66	0.72
	Real + Synthetic	LR	80.1% (+9.9%)	0.76	0.83
	Real Only	GBM	74.50%	0.71	0.76
	Real + Synthetic	GBM	83.7% (+9.2%)	0.81	0.86

Random Forest and Gradient Boosting achieve the highest accuracy improvement (~10%) on small datasets. These models handle non-linear relationships and feature interactions well, making them more robust to augmented data. Although SVM and KNN also improve, their gains are slightly lower (~5-9%) in the Iris and Heart Disease datasets. KNN's reliance on distance-based learning may make it more sensitive to small variations in synthetic data.

To investigate whether the adapted CTGAN is applicable to large datasets, initial experiments were conducted on two benchmark datasets (i.e., Bank Marketing and Loan Default Prediction). Table 7 presents the properties of the datasets, including the number of synthetic samples added to the dataset. For the Marketing dataset, CTGAN generated 34,500 synthetic samples to ensure the 'No' samples were approximately equal to the 'Yes' samples. As for the Loan Default Prediction dataset, both classes are balanced with 85,000 samples in each class. Data in Table 8 shows the classification outcomes (averaged across the two datasets). The F1-score demonstrates an improvement of approximately 9–10%, primarily attributed to enhanced recall for the minority class. This performance gain is accompanied by a modest increase in overall accuracy (~1.5%), indicating improved generalisation capabilities of the models. Among the evaluated classifiers, GBM and RF yield the best performance, likely due to their robustness in handling the variability introduced by synthetic data.

Table 7

Properties of Large Datasets

Dataset	Samples	Numerical Features	Categorical Features	# Synthetic Samples
Marketing	45,211	17	8	34,500
Loan Default Prediction	100,000	14	9	70,000

Table 8

Classification (Average) of Large Datasets

Classifier	Accuracy (Real Data)	Accuracy (Real + Synthetic)	F1-Score (Real)	F1-Score (Real + Synthetic)
RF	87.0%	88.5% (+1.5%)	0.72	0.81 (+9%)
KNN	82.5%	84.0% (+1.5%)	0.65	0.75 (+10%)
SVM	85.2%	86.8% (+1.6%)	0.68	0.78 (+10%)
GBM	89.5%	91.0% (+1.5%)	0.74	0.83 (+9%)

The utilized Bank Marketing dataset contains a minority class that comprises customers who subscribed to a term deposit (approximately 11% of the dataset), while the majority class includes those who did not. The application of CTGAN significantly enhanced the model's ability to predict term deposit subscriptions, yielding an 8 – 10% improvement in recall and F1-score. Additionally, CTGAN reduced the model's bias toward predicting non-subscription outcomes. On the other hand, the Loan Default Prediction includes data on loan defaulters, which represent the minority class (approximately 15% of the dataset), and non-defaulters constitute the majority. The integration of CTGAN led to a notable improvement in the detection of defaulters, with recall and F1-score increasing by approximately 9 – 10%. This enhancement reduces false negatives, indicating fewer missed default cases. However, the improved sensitivity was accompanied by a slight increase in false positives, suggesting a trade-off in predictive precision. Details on the classification outcomes for both the Bank Marketing and Loan Default Prediction are depicted in Table 9.

Table 9

Detailed Classification Results of Large Datasets

Classifier	Dataset	Data Type	Precision	Recall
RF	Bank Marketing	Real	0.70	0.74
		Real + Synthetic	0.79	0.83
	Loan Default Prediction	Real	0.75	0.69
		Real + Synthetic	0.85	0.78
KNN	Bank Marketing	Real	0.63	0.67
		Real + Synthetic	0.73	0.77
	Loan Default Prediction	Real	0.68	0.62
		Real + Synthetic	0.78	0.72
SVM	Bank Marketing	Real	0.66	0.70
		Real + Synthetic	0.77	0.79
	Loan Default Prediction	Real	0.70	0.66
		Real + Synthetic	0.80	0.76
GBM	Bank Marketing	Real	0.72	0.76
		Real + Synthetic	0.82	0.85
	Loan Default Prediction	Real	0.78	0.70
		Real + Synthetic	0.87	0.79

The GBM consistently achieved the highest performance, reaching 91.0% accuracy and an F1-score of 0.83 with the enhanced dataset. Precision and recall values also increased for both Bank Marketing and Loan Default Prediction datasets, reflecting a better balance between true positives and false positives. These improvements suggest that data augmentation with synthetic samples can enhance model robustness and generalisation, especially in classification tasks with limited or imbalanced real-world data.

CONCLUSION

Data augmentation is crucial for enabling effective learning from small datasets in data analytics because it directly addresses the inherent limitations of limited data. Small datasets are prone to overfitting, where models memorize training examples rather than learning generalizable patterns; augmentation overcomes this by introducing variations in the data, promoting better generalisation. Additionally, by simulating real-world noise and distortions, augmentation enhances model robustness, making it more resilient to data imperfections. Finally, in addressing data imbalance, augmentation can provide synthetic data for under-represented classes, assuring balanced training and better performance. In this study, the CTGAN effectively augments small datasets as the generated data is shown to have statistical similarity (low Wasserstein distance, low JSD values) with the real data. This leads to the ability to provide more data for a machine learning model to learn. When more data is fed, the deployed machine learning model performs better (higher accuracy, F1-score, AUC-ROC). Further investigation on the effectiveness of CTGAN revealed that the network is also a promising mechanism for large datasets with imbalanced classes. The addition of synthetic data improved accuracy and F1-score across all classifiers, where the GBM consistently achieved the highest performance.

This study could be further improved by investigating adaptive augmentation, a dynamic approach where augmentation strategies evolve in response to the model's learning progress. Unlike current techniques, including CTGAN, which apply static transformations regardless of model feedback, adaptive augmentation can tailor the' type, intensity, or frequency of augmentations based on real-time model performance metrics such as loss curves, prediction confidence, or class-wise error rates. For instance, a model struggling with underrepresented classes could trigger the generation of more synthetic examples specifically for those classes, or apply stronger transformations to overfit-prone data regions. By incorporating a feedback loop, adaptive augmentation holds the potential to enhance generalisation more effectively, reduce overfitting, and improve robustness across diverse and imbalanced datasets.

ACKNOWLEDGMENT

This research was supported by the Ministry of Higher Education (MOHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2022/ICT02/UUM/02/1).

REFERENCES

Ahsan, M., Gomes, R., Denton, A., Alom, M. Z., Mohammad, N., & Mahmood, M. (2025). Hybrid oversampling technique for imbalanced pattern recognition: Enhancing performance with borderline synthetic minority oversampling and generative adversarial networks (BSGAN). *Machine Learning with Applications, 20*, Article 100637. https://doi.org/10.1016/j.mlwa.2024.100637

Biau, G., Sangnier, M., & Tanielian, U. (2021). Some theoretical insights into wasserstein gans. *Journal of Machine Learning Research, 22*. https://doi.org/10.48550/arXiv.2006.02682

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*. https://doi.org/10.1613/jair.953

Chen, J., Yang, Z., & Yang, D. (2020). MixText: *Linguistically-informed interpolation of hidden space for semi-supervised text classification*. Proceedings of the Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.194

Dua, M., Joshi, S., & Dua, S. (2023). Data augmentation based novel approach to automatic speaker verification system. *E-Prime - Advances in Electrical Engineering, Electronics and Energy, 6*. https://doi.org/10.1016/j.prime.2023.100346

Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A comprehensive survey of generative adversarial networks (GANs) in cybersecurity intrusion detection. *IEEE Access, 11*. https://doi.org/10.1109/ACCESS.2023.3296707

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). *Understanding back-translation at scale.* Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018. https://doi.org/10.18653/v1/d18-1045

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets.* Learning from Imbalanced Data Sets. https://doi.org/10.1007/978-3-319-98074-4

Fidon, L., Ourselin, S., & Vercauteren, T. (2021). *Generalized wasserstein dice score, distributionally robust deep learning, and ranger for brain tumor segmentation: BraTS 2020 challenge.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12659 LNCS. https://doi.org/10.1007/978-3-030-72087-2_18

Goodfellow et al. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11). https://doi.org/10.1145/3422622

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2023). A review on generative adversarial networks: algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering, 35*(4). https://doi.org/10.1109/TKDE.2021.3130191

Habibi, O., Chemmakha, M., & Lazaar, M. (2023). Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. *Engineering Applications of Artificial Intelligence, 118*. https://doi.org/10.1016/j.engappai.2022.105669

Hou, L., Cao, Q., Shen, H., Pan, S., Li, X., & Cheng, X. (2022). *Conditional GANs with auxiliary discriminative classifier.* Proceedings of Machine Learning Research, 162. https://doi.org/10.48550/arXiv.1610.09585

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., & Gómez-Canaval, S. (2023). Data Augmentation techniques in time series domain: A survey and taxonomy. *Neural Computing and Applications, 35*(14). https://doi.org/10.1007/s00521-023-08459-3

Isensee, F., Jäger, P. F., Full, P. M., Vollmuth, P., & Maier-Hein, K. H. (2021). *nnU-Net for brain tumor segmentation.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12659 LNCS. https://doi.org/10.1007/978-3-030-72087-2_11

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications, 244*. https://doi.org/10.1016/j.eswa.2023.122778

Khan, A. R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S., & Mehmood, Z. (2021). Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. *Microscopy Research and Technique, 84*(7). https://doi.org/10.1002/jemt.23694

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning, 12*(4). https://doi.org/10.1561/2200000056

Kobayashi, S. (2018). *Contextual augmentation: Data augmentation bywords with paradigmatic relations.* NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2. https://doi.org/10.18653/v1/n18-2072

Lyu, Y., & Tian, X. (2025). MWG-UNet++: Hybrid transformer U-Net model for brain tumor segmentation in MRI scans. *Bioengineering, 12*(2), 140. https://doi.org/10.3390/bioengineering12020140

Ma, H., Geng, M., Wang, F., Zheng, W., Ai, Y., & Zhang, W. (2024). Data augmentation of a corrosion dataset for defect growth prediction of pipelines using conditional tabular generative adversarial networks. *Materials, 17*(5). https://doi.org/10.3390/ma17051142

Majeed, A., & Hwang, S. O. (2023). CTGAN-MOS: Conditional generative adversarial network based minority-class-augmented oversampling scheme for imbalanced problems. *IEEE Access, 11*. https://doi.org/10.1109/ACCESS.2023.3303509

Park et al. (2019). *Specaugment: A simple data augmentation method for automatic speech recognition.* Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-September. https://doi.org/10.21437/Interspeech.2019-2680

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). *The synthetic data vault.* Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016. https://doi. org/10.1109/DSAA.2016.49

Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., & Tao, D. (2021). *SynFace: Face recognition with synthetic data.* Proceedings of the IEEE International Conference on Computer Vision. https://doi. org/10.1109/ICCV48922.2021.01070

Sainin, M. S., Alfred, R., & Ahmad, F. (2021). Ensemble meta classifier with sampling and feature selection for data with imbalance multiclass problem. *Journal of Information and Communication Technology, 20*(2), 103–133. https://doi.org/10.32890/jict2021.20.2.1

Sennrich, R., Haddow, B., & Birch, A. (2016). *Improving neural machine translation models with monolingual data.* 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 1. https://doi.org/10.18653/v1/p16-1009

Strelcenia, E., & Prakoonwit, S. (2023). A survey on GAN techniques for data augmentation to address the imbalanced data issues in credit card fraud detection. *Machine Learning and Knowledge Extraction, 5*(1). https://doi.org/10.3390/make5010019

Tavor et al. (2020). *Do not have enough data? Deep learning to the rescue!* AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v34i05.6233

Tsalera, E., Papadakis, A., Pagiatakis, G., & Samarakou, M. (2025). Impact evaluation of sound dataset augmentation and synthetic generation upon classification accuracy. *Journal of Sensor and Actuator Networks, 14*(5), 91. https://doi.org/10.3390/jsan14050091

Uchitomi, H., Ming, X., Zhao, C., Ogata, T., & Miyake, Y. (2023). Classification of mild Parkinson's disease: Data augmentation of time-series gait data obtained via inertial measurement units. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-39862-4

Wei, J., & Zou, K. (2019). EDA: *Easy data augmentation techniques for boosting performance on text classification tasks.* Proceedings of the EMNLP-IJCNLP 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. https://doi.org/10.18653/v1/d19-1670

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information (Switzerland), 14*(1). https://doi.org/10.3390/info14010054

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems, 32*. https://doi.org/ 10.48550/arXiv.1907.00503

Zarandah, Q. M. M., Daud, S. M., & Abu-Naser, S. S. (2023). Spectrogram flipping: A new technique for audio augmentation. *Journal of Theoretical and Applied Information Technology, 101*(11). http://www.jatit.org/volumes/Vol101No11/26Vol101No11.pdf

Zhang et al. (2023). GAN-based one-dimensional medical data augmentation. *Soft Computing, 27*(15). https://doi. org/10.1007/s00500-023-08345-z

Zhao, X., & Guan, S. (2023). CTCN: A novel credit card fraud detection method based on conditional tabular generative adversarial networks and temporal convolutional network. *PeerJ Computer Science, 9.* https://doi.org/10.7717/PEERJ-CS.1634