Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-024-02120-5

ORIGINAL RESEARCH

A synergistic fusion of shallow and deep generative model to enhance machine learning efficacy and classification performance in data-scarce environments

Khursheed Ahmad Bhat[1] · **Shabir Ahmad Sofi**[1]

Received: 29 April 2024 / Accepted: 1 August 2024

Abstract Data-constrained environments present a significant challenge to the effectiveness of machine learning and deep learning algorithms. The performance of these algorithms is inherently dependent on the quality and quantity of the training data they are exposed to. When training data is insufficient, the efficacy of downstream learning tasks is markedly diminished. The constrained availability and scarcity of data result from diverse intrinsic factors, encompassing data regulations, privacy concerns, the confidential nature of data, and the inherent rarity of data of interest in critical real-world applications. To tackle this, researchers have come up with the concept of synthetic data to provide a low-cost, easily available and secure alternative. Synthetic data serves to bolster the robustness of model learning within real-world contexts, addressing the formidable challenge posed by data scarcity. This scarcity leads to class imbalance problem and insufficient representation of data very often encountered in real world datasets. Popular data generation strategies involve increasing the representation of minority class instances through the generation of synthetic examples. The existing data generation techniques aim to expand datasets for balancing, yet they frequently fall short in achieving satisfactory sample diversity. This paper examines the potential of shallow interpolation based data generation technique to capture the local dynamics for minority balancing and deep generative modeling based generative adversarial networks (GANs) for global distribution estimation for augmentation the smaller datasets. This paper presents a hybrid amalgamation approach for tabular data involving mixed type attributes and pays special attention to data imbalance and insufficient data problems. We named this approach as hybrid data balancing and augmentation approach on mixed tabular data (HDBA-MTD) tailored for synthesizing samples on underrepresented target labels (skewed class labels) and insufficient data instances. This approach exhibits the potential to restore dataset balance, address bias in the dataset, mitigate over-fitting issues, enhance training data diversity, thereby pays special attention towards the downstream classification and generalization performance in data rarity dilemmas. Experiments are conducted on benchmark datasets to validate the feasibility of the proposed model in realistic scenarios. The evaluation and analysis of experimental procedures demonstrate favorable comparisons with other existing synthetic data generation techniques.

Keywords Insufficient data · SMOTE · GANs · Synthetic data · Minority class · Data balancing and augmentation

1 Introduction

The revolutionary advancements in science and technology over the past decade have facilitated an unprecedented expansion and availability of raw data [1, 2]. This surge has propelled remarkable progress in the fields of artificial intelligence (AI), machine learning (ML), and deep learning (DL). ML and DL models have played a pivotal role in seamlessly addressing various real-world problems, including medical diagnosis [3], fault detection [4], failure prediction [5], online fraud [6], survival predictions [7], pattern recognition [8],anomaly detection, and network intrusion [9], sentiment analysis [10] among others. Beyond the substantial

✉ Khursheed Ahmad Bhat
k.a.bhat_002pha2019@nitsri.ac.in

Shabir Ahmad Sofi
shabir@nitsri.ac.in

[1] National Institute of Technology, Srinagar, JK, India

Published online: 09 August 2024

computational power, the exceptional performance in prediction and classification tasks can be attributed to the abundance of data accessible to learning algorithms. These algorithms have achieved significant milestones in prediction, classification, and detection tasks [11]. Exploration into machine learning has unveiled its capacity to unlock novel applications across diverse disciplines.

The data scarcity and class imbalance present major hurdles for classification and prediction problems in machine learning. Insufficient data poses significant challenges in ML and DL, limiting data analytics and insights across diverse research domains [11]. The reliability of ML algorithms for prediction and classification hinges on the dataset's nature. While an ideally balanced and diverse (adequately representative) dataset with even class distributions is optimal, such datasets are rare, with most exhibiting significant imbalance skewed towards the majority group. The limited availability and scarcity of data stem from various intrinsic reasons, such as data regulations, privacy concerns, proprietary data, confidential and sensitive information, under representation of data of interest, and the inherent rarity of certain data species [12]. It is well-established that ML and DL algorithms exhibit optimal learning efficiency in the presence of qualitative, diverse and balanced datasets [13]. Therefore, considering the challenges associated with limited data, a viable solution, characterized by being low-cost, privacy-preserving, non-conflicting, and readily available, involves exploring the avenue of synthetic data generation [14]. The goal is to address the challenge of data imbalances and data scarcity by leveraging synthetic data. In various real-world applications, data scarcity raises serious concerns, reducing the model's prediction performance. Data distribution imbalance reduces the models' classification accuracy used in practical cases.

In the same vein that AI endeavors to create machines that emulate human thought processes, synthetic data aims to generate artificial datasets mirroring the characteristics of real-world data. Within the realm of synthetic data, various methods and techniques for generating data exist. Synthetic data fundamentally serves as an oversampling technique commonly employed for augmentation [14]. The primary objective of data augmentation is to enhance the volume, quality, diversity of training data which helps to mitigate prediction and misclassification problems owing to data scarcity problem. In the context of class imbalance, we observe that it is common to mitigate the imbalance issue by creating synthetic samples of the minority class. Data augmentation also helps to over-fitting problem by acting as an implicit regularization technique [15]. The key contributions can be outlined as follows:

1. The presented approach, hybrid data balancing and augmentation approach on mixed tabular data (HDBA-MTD), employs a hybrid synthetic minority oversampling technique (SMOTE) based data balancing technique to effectively address class imbalance problem followed by conditional tabular generative adversarial network (CTGAN) for data augmentation/enhancement in multimodal and mixed type tabular datasets for boosting the downstream ML efficacy.
2. The current research utilizes a strategy involving both label balancing and data augmentation through synthetic data paradigm. This approach is adopted to effectively learn the local data and global data distribution.
3. This research work comprehensively examines structured datasets, encompassing heterogenous data types, non-gaussian and multimodal distributions, as well as highly imbalanced categorical columns. The exploration includes theoretical and experimental aspects of debiasing, label balancing and data augmentation for improved classification performance.
4. To mitigate discrepancies in the implementation and evaluation details, the investigation approach uses model agnostic metrics for evaluating data quality and performance of classification tasks.

The rest of the paper is organized into the following sections. Section II covers the background and related works in the field. Section III details the methodologies adopted to achieve the goal of debiased balanced datasets for classification and generalization. It also details about the presented approach employed in this work. This section also focuses on the experimental setup, implementation details including evaluation metrics employed for model prediction and data evaluation. The results in the form of research questions (RQs) based on the experimental procedure is presented in Section IV. Section V highlights the discussion and analysis on the results obtained in section IV. Finally, Section VI concludes with a summary of this study and insights about future research.

2 Background & related works

The impetus to do research on synthetic data generation stems from the potential to tackle real-world data scarcity challenges. The availability of large datasets is critical and necessary for ML and DL models. Different ML and DL models will not be able to learn and build reliable models unless there are many visuals in datasets. Unfortunately, data is largely imbalanced and limited in majority of real-world domains, which inhibits efficient and effective learning. Numerous real-world applications are known to suffer from the problem of class imbalance and data scarcity problem [16]. Data distribution imbalance, reduces the models' classification accuracy used in practical cases. Many classification applications, including disease detection, anomaly detection, fraud detection, intrusion detection, and failure prediction etc., inevitably face the issue of class imbalances. Using synthetic data to address class

label disparity can be a crucial solution for these application areas, leading to improvements in downstream classification tasks [11, 16].

Synthetic data generation encompasses various techniques, including geometric transformations, resampling methods, and data synthesis methods [11]. Classical transformation methods, mostly employed in image domain involve operations like scaling, rotating, and flipping [17]. The resampling method includes both random under sampling (RUS), which handles imbalance by randomly truncating majority class samples, and random oversampling (ROS), which addresses imbalance by randomly increasing minority class samples to restore balance between the two classes [18]. The transformations are simple data augmentation techniques specific to the image domain, may contain duplicates and do not intelligently model the diversity of data samples. Similarly, classical resampling methods are uninformed and blind in nature, it may create more duplicates and may not be true representative at all, therefore, may lead to over-fitting [19]. The data synthesis approach involves generating artificial samples to inflate the size of the datasets. This includes distance-based techniques like synthetic minority over-sampling technique (SMOTE) and its variants, as well as density estimation-based deep generative modeling techniques, where generative adversarial networks (GANs) are particularly popular and state-of-the-art compared to other methods [20, 21]. The synthetic data generation is a data-oriented approach which work by affecting data space directly and attempting to lower the imbalance ratio between classes and enrich the overall representation of data instances. Data oriented approaches offer greater flexibility and are easy to implement, as they can be applied with any classification algorithm [22, 23]. The SMOTE stands out as one of the most commonly utilized oversampling methods [24], which generates additional samples for the minority class through linear interpolation between neighboring minority cases. While lesser-known distance-based SMOTE variations exist in the literature, a study conducted by [25] observes that SMOTE is the go-to method of data generation for tackling class imbalance problem. The new variations relying on SMOTE may not necessarily surpass their forerunner in terms of performance. Currently, a multitude of real-world applications deal with non-linear, complex data that includes both continuous and discrete variables. The different variables information content often exhibits heterogeneity, with some variables carrying more predictive information than others. In such scenarios, computing distances based on SMOTE alone to identify neighboring minority cases becomes challenging [25, 26]. On the contrary, GANs have the capability to learn complex, non-linear, heterogeneous, high-dimensional distributions, potentially making them more effective for oversampling [20, 21].

In recent years, there has been a notable increase in the interest surrounding oversampling using GANs in the machine learning domain. This is attributed to their capability to model intricate and high-dimensional distributions, particularly in situations of class imbalance. Consequently, a surge in research endeavors has emerged, aiming to develop new architectures, techniques, and applications for GANs [27–31]. Notably, GANs have found extensive use in synthesizing images for computer vision tasks [32–35] [36]. As an augmentation scheme, they have demonstrated significant performance enhancements, as emphasized in a recent survey by the authors in research work [33]. However, the scenario differs in the context of tabular data modeling. Structured data poses heightened challenges compared to unstructured data (such as images and text) due to the amalgamation of categorical and numerical columns, absence of contextual and local information inherent in image and sequence data, and the skewed distribution of categorical columns [38]. This means that only handful of studies use GANs for oversampling structured tabular data. The recent studies conducted by [26, 39–43] has used smote in combination with deep generative modeling techniques. SMOTE is the common kick starter approach in these studies, however, they are using vanilla GANs which are not suitable in the context of mixed tabular data [44]. The study conducted by [26] bears notable similarities to our approach. However, it predominantly focuses on evaluating data synthesis methodologies concerning disease prediction rather than testing on highly imbalanced datasets. The study by [45]employ a conventional GAN to train on minority class observations, subsequently utilizing the trained generator for oversampling. Their study involves a comparison between GAN, SMOTE, and no oversampling using a single classification algorithm on a credit card fraud dataset with exclusively numerical variables and not testing on categorical columns. The research work by [46] introduced an oversampling approach based on Conditional Generative Adversarial Networks (cGAN). The conditional approach allowed to tackle the mode collapse problem of GAN architectures. However, the study utilizes the default GAN loss function, known to encounter stability and convergence issues [47]. Furthermore, the evaluation of their method extends to both real-world and artificial datasets, exclusively comprising continuous attributes only. While GANs have demonstrated significant potential in modeling intricate data distributions, the exploration and the limitations of the architectures remains an ongoing research challenge. Investigating these limitations is crucial, as deeper architectures have the potential to capture more intricate relationships within data and better approximate the original data distribution.

With reference to tabular data, tabular generative adversarial network (TGAN) was developed by the authors in [48]. The architecture uses long-short-term-memory (LSTM) model and gaussian mixture model (GMM) which was soon replaced by the conditional tabular generative adversarial network (CTGAN) developed by the same authors in the research work [44]. CTGAN is based on more sophisticated

fully connected networks unlike LSTM model in TGAN and variational gaussian mixture model which automatically detects the number of modes in continuous columns instead of GMM [49]. CTGAN is acknowledged as the leading and state-of-the-art GAN based modeling technique for handling structured tabular data [50, 51]. The CTGAN architecture by [44] opts for conventional fully-connected layers for generator and discriminator. Generator and discriminator are two neural networks that form the core of any GAN architecture including the adapted CTGAN in this research work. These networks engage in an adversarial competition to produce synthetic yet realistic data. Unlike previous architectures utilized in prior studies, this architecture is well-suited to model heterogeneous columns and mixed-type attributes of tabular data. Meanwhile, for numerical columns, it embraces an approach based on a gaussian mixture model. Additionally, CTGAN replaces the default loss function for improved stability and convergence [44, 51]

The literature on addressing data scarcity dilemmas is continuously evolving. Before the advent of deep generative models like GANs, classical transformations and primitive resampling techniques were used to inflate dataset sizes and restore target class equilibrium [52]. These primitive techniques have significant limitations, as they are uninformed and lack sophistication. Blindly increasing the size of data can negatively affect model learning and lead to data fitting problems [11].

To address these issues, the authors in [24] introduced the data synthesis approach in the form of SMOTE, which creates synthetic data points between existing real data points. SMOTE helps to mitigate data scarcity and is statistical in nature. However, it does not perform well with large datasets and high-dimensional data [11].

With the advent of generative models, which excel at handling large and high-dimensional datasets and can model entire data distributions, techniques like SMOTE have become less prominent. However, GANs require a considerable amount of data to effectively model and synthesize new data from the latent distribution. They struggle to learn efficiently with very few data samples [11, 53].

In the context of larger datasets and big data, GANs outperform distance-based shallow techniques and represent the state-of-the-art [37, 54]. However, complex datasets with data scarcity issues remain challenging. Single augmentation techniques are often insufficient, leading to increased interest in hybrid approaches that combine multiple methods to better address these challenges.

The presented approach in this research draws its inspiration from the works conducted by [26, 39, 40, 43]. Each of these studies have undertaken a hybrid mechanism to deal with data scarcity and class imbalanced problem. Our work is different in many aspects from the existing ones in the literature to prove its novelty by enough margin. The existing works have focused on only increasing the minority samples through the combination of SMOTE and GAN techniques. The authors in [40] uses minority samples from SMOTE to train the GAN model and generate new minority class instances, while the authors in [39] uses minority class samples as an input to the generator module of the GAN instead of latent space noise vector. Their reliance on unconditional variants of GAN architectures represents their most significant limitation. The work conducted by [26] uses many variants of smote with combination of conditional tabular GAN but is mainly focused on the evaluation of data generation in disease prediction and also the datasets are comparably balanced in the first place. The key findings from the most significant related literature are presented in Table 1.

We differ by a significant margin by changing the methodology of the data pipeline. The shallow distance-based minority class balancing in data is crucial because deep generative models like GANs cannot be directly applied to highly imbalanced data. In situations of extreme class imbalance and data scarcity, the available data may not be fully representative for effective learning by deep generative models. The synthetic data augmentation, coupled with shallow SMOTE and based on deep generative density estimation modeling, plays a crucial role in enriching the diversity of data samples and expanding the overall dataset, leading to improved prediction and classification outcomes. By initializing the generator module of CTGAN with a noise vector (random input), unlike the methods used in [39, 40], our approach potentially has an advantage in preserving data privacy, as the real data is not exposed to the generator during the synthesis of data samples.

We have artificially created an extreme class imbalance with imbalance ratio (IR) > 20 [22] in order to test the efficacy of this intelligent oversampling approach. The novelty of this research lies in the presented approach, the hybrid data balancing and augmentation approach on mixed tabular data (HDBA-MTD), employs a SMOTE based synthetic data balancing technique to effectively address class imbalance problem followed by CTGAN for data augmentation/enhancement in multimodal and mixed type tabular datasets for boosting the downstream ML efficacy. Our findings contribute to improved generalization to unseen data and enhance the performance of downstream tasks in data rarity dilemmas.

3 Methodology

This work uses fusion of two different informed oversampling artificial data generation techniques. The methodology is a hybrid scheme consisting of data balancing and data augmentation module. The study introduces an innovative methodology characterized by its customized sampling conditions and meticulous control on sample generation,

Table 1 Different data augmentation techniques employed in literature for handling data scarcity problems

Technique	Objective	Domain	Findings	Ref
Geometric transformations	Data augmentation	Image data	Transformations operations include scaling, rotating, and flipping, predominantly used in the image domain for increasing the size of image datasets	[17]
Resampling methods (ROS and RUS)	Class imbalance handling	Multi domain	These sampling methods are primitive approaches to tackle disparity of data samples with respect to target class. RUS reduces majority class samples while ROS increases minority class samples. These techniques are blind and uninformed in nature and cannot model the true distribution of data	[18]
SMOTE	Class imbalance handling	Multi domain	SMOTE extends random oversampling by generating synthetic minority class samples. These are generated by choosing a random minority case and creating a linear combination with a neighboring minority case, randomly selected from the k-nearest neighbors. SMOTE is the most popular synthetic data generation based statistical technique which works well in low dimensional data but face problems in modeling complex domain data, high dimensional data, big data and mixed type tabular data	[24]
SMOTE	Class Imbalance handling	Tabular data	This study addresses the imbalanced nature of NASA datasets by utilizing oversampling through SMOTE, a shallow synthetic data generation method based on distance. However, it's important to note that the study does not compare SMOTE with any other data generation approach, which represents a significant limitation	[4]
GAN	Data augmentation	Image data	Within the realm of healthcare and disease diagnosis, researchers are increasingly turning to the versatile capabilities of the vanilla GAN architecture for data augmentation, particularly when working with image datasets. Unlike conventional resampling and transformation techniques, which often struggle to accurately represent real-world data, these studies demonstrate the GAN model's unique ability to synthesize data that closely mirrors the actual distribution. This growing body of research, among others exploring GAN applications in the image domain, serves as a resounding testament to the remarkable success of GANs in revolutionizing image-based data domain	[32–34, 36]

Table 1 (continued)

Technique	Objective	Domain	Findings	Ref
GAN	Class imbalance handling	Tabular data (numerical columns only)	This research work employs a conventional GAN to train on minority class observations, subsequently utilizing the trained generator for oversampling. Their study involves a comparison between GAN, SMOTE, and no oversampling using a single classification algorithm on a credit card fraud dataset with exclusively numerical variables and not testing on categorical columns	[45]
N-GAN	Synthetic data generation	Tabular (anomaly detection)	This paper centers on estimating the global distribution of malicious samples to determine the threshold for relevant data. The results, tested on a single intrusion detection dataset, support the use of GANs for anomaly detection	[54]
Conditional GAN	Class imbalance handling	Tabular data (numerical columns only)	This work introduced an oversampling approach based on Conditional Generative Adversarial Networks (cGAN). The conditional approach allowed to tackle the mode collapse problem of GAN architectures. However, the study utilizes the default GAN loss function, known to encounter stability and convergence issues	[46]
Hybrid (Oversampling and Under sampling)	Class imbalance handling	Image domain	This study aims to improve the classification accuracy of hyperspectral images, particularly addressing the challenge of minority class imbalance. It employs a hybrid resampling strategy, combining SMOTE-based oversampling to generate more instances of minority classes with under-sampling of majority class samples by replacing them with cluster centroids in a multiclass scenario. It's essential to acknowledge that under-sampling could result in the loss of valuable information	[52]
Hybrid (Under sampling + GAN)	Imbalanced data handling	Tabular data	The paper introduces a novel method that combines under-sampling and over-sampling techniques with association rule learning and GANs to address the challenges of imbalanced classification. This work does not explain how it compensates for the information loss resulting from under-sampling. Additionally, it employs TGAN, which has since been superseded by the state-of-the-art CTGAN architecture, making it less effective at modeling multiple modes in continuous columns	[19]

Table 1 (continued)

Technique	Objective	Domain	Findings	Ref
Hybrid Data synthesis (SMOTE + GAN)	Data augmentation	Tabular data	The aim is to explore synthetic data modeling in the absence of real data. It is focused on the evaluation of data generation in disease prediction and also the datasets are comparably balanced in the first place	[26]
Hybrid Data synthesis (SMOTE + GAN)	Class imbalance handling	Tabular data (numerical columns)	The authors in this work uses minority class samples as an input to the generator module of the GAN instead of latent space noise vector. Their reliance on unconditional variants of GAN architectures represents their most significant limitation	[39]
Hybrid (SMOTE + GAN)	Class imbalance handling	Tabular data Internet of Things (IoT)	This study uses minority samples from SMOTE to train the GAN model and generate new instances for the minority class. Their dependency on using artificially generated samples to train the GAN network instead of real data samples represents a major limitation, in addition to employing a vanilla GAN architecture	[40]
Hybrid (GAN + SMOTE)	Class imbalance handling	Tabular data (petrography)	This work discusses the use of GAN architecture first for synthesizing artificial samples followed by minority class augmentation through vanilla GAN. Furthermore, the dependency on unconditional variants of GAN architectures stands as their most significant limitation, as it lacks control over the generation of sample data	[41]
Hybrid (SMOTE + GAN)	Class imbalance handling	Tabular Financial data	This study investigates the applicability of combining distance-based sample generation techniques with distribution estimation data generation methods in the context of financial datasets for detecting and identifying fraudulent transactions. However, this approach faces challenges in adequately handling the multiple modes of continuous columns and the highly skewed nature of discrete columns	[43]
TGAN	Modeling tabular data	Tabular datasets	With the rising popularity of GANs with respect to image and textual domain, a special GAN dedicated to tabular datasets was developed and was called as TGAN. This architecture uses, gaussian mixture model (GMM) with fixed modes to model continuous columns. This architecture was not used significantly because it was soon replaced by the state-of-the-art CTGAN architecture which is based on more sophisticated fully connected network architecture and variational gaussian mixture model which automatically detects the number of modes in continuous columns	[48]

Table 1 (continued)

Technique	Objective	Domain	Findings	Ref
CTGAN	Modeling tabular data and Class imbalance handling	Different tabular datasets	Each of these studies employed CTGAN for modeling tabular data. CTGAN can handle the complexity of tabular datasets more efficiently than other GAN variants discussed in the literature. However, like other deep learning-based models, CTGAN also struggles with acute data scarcity and may not sufficiently learn the entire representation of data samples. Therefore, on smaller datasets, CTGAN may not guarantee optimal performance	[44, 50, 51]
HDBA-MTD (SMOTE + CTGAN)	Class imbalance handling and data augmentation	Tabular data (numerical and continuous columns)	A novel approach, hybrid data balancing and augmentation approach on mixed tabular data (HDBA-MTD), combines SMOTE for data balancing and CTGAN for data augmentation. It is specifically designed for mixed tabular data with both continuous and discrete columns and a highly skewed target class tilted in favor of the majority class. The synergistic fusion of shallow and deep model used in this work efficiently complements each other in data scarcity environments. This approach increases dataset samples, enhances diversity for effective generalization to unseen data. This approach thereby enhances the downstream classification performance of machine learning tasks compared to existing methods	This work

rendering its applicability to diverse real-world data within the context of classification tasks. Utilizing both traditional data generation methods and cutting-edge deep learning techniques, the aim is to produce synthetic data samples to address class imbalance and expand smaller datasets, thus rendering them more suitable for downstream machine learning tasks. This approach primarily seeks to improve model performance by mitigating bias introduced by skewed data and augmenting data volume to facilitate better generalization. The methodology of the research work is presented with the help of algorithm 1 below.

Algorithm 1 Methodology employed to generate a synthetically augmented and balanced dataset for increasing the diversity of data samples using the HDBA-MTD approach

Input: Real dataset D_{real} (small and imbalanced). Target class labels Y_M and Y_m where Y_M denotes majority class label and Y_m denotes minority class label. $|Y_m| < |Y_M|$ implies that input dataset is class imbalanced.
Module 1: Class balancing using SMOTE (shallow data synthesis based on nearest neighbours)
Output: Class balanced dataset D_{bal} such that $Y_M = Y_m$
Require: Original dataset D_{real} $|Y_M|$ and $|Y_m|$ denote the number of majority class samples and minority class samples respectively
for each datapoint in minority class y_m **do**
    nearest neighbours = find k nearest neighbours (datapoint, Y_m, k=5)
    ▷ k= 5 is default hyperparameter of SMOTE
    for each neighbour in nearest neighbours **do**
        new minority sample = create synthetic data (datapoint, neighbour)
        Obtain balanced dataset D_{bal} where class labels are balanced
    end for
end for
Module 2: Synthetic augmented dataset using CTGAN (deep generative modelling data synthesis based on estimating the entire probability distribution of data samples)
Output: Synthetic augmented dataset D_{synaug} obtained by employing CTGAN on D_{bal}
Require: Balanced dataset D_{bal} obtained from ***Module 1***
Initialize the hyper parameters of Generator (G) and Discriminator (D) of CTGAN with default settings (no. of epochs, batch_size, learning rate)
for epochs = 1 to no. of training epochs **do**
    for each mini_batch **do**
        sample fake data samples from $p_g(z)$ using generator G ▷ $p_g(z)$ = probability ▷ distribution of fake data samples synthesised with generator on random input z

        sample real data samples from data distribution $p_{data}(x)$ ▷ $p_{data}(x)$ = probability ▷ distribution of real data samples

        Compute discriminator loss for real and synthetic samples
        Update the discriminator using backpropagation
    end for
    Compute the generator loss using discriminator networks output
    Update the generator network using back propagation
end for
for samples = 1 to N_{syn} **do** ▷ N_{syn} denotes the number of samples to generate ▷ based on balanced dataset D_{bal} such that $N_{syn} = D_{bal}$
    Sample the number of data instances required to create synthetic dataset D_{synaug}
end for
Final Output: $D_{real\_syn} = D_{real} + D_{synaug}$ ▷ Resultant diverse and enriched dataset obtained by employing ***Module 1*** and ***Module 2*** of this algorithm. D_{real} and D_{synaug} are merged to yield final dataset D_{real_syn}

Table 2 Commonly used notations and their respective explanations

Symbol	Details	Symbol	Details
D_{real}	Original dataset	D_{bal}	SMOTE balanced dataset
D_{synaug}	Synthetic GAN dataset	D_{real_syn}	Final merged dataset
Y_M	Majority class	Y_m	Minority class
x_i	Random minority sample	x_j	Randomly selected from k nearest sample
▷	comment		

This research focuses on addressing a binary classification problem in n-dimensional real space, using a real-world dataset 'D_{real}' comprising input features and target classes. D_{real} contains a mix of continuous and discrete attributes, with each row representing a sample containing all data. The goal is to achieve balanced learning between two classes' 'Y_M' and 'Y_m', where Y_M represents the dominant class and Y_m the dominated class. To balance the dataset, SMOTE is employed to generate synthetic data points based on local information, resulting in the dataset 'D_{bal}'. The balanced dataset 'D_{bal}' is then augmented with synthetic data generated by CTGAN, resulting in the dataset 'D_{synaug}'. Finally, D_{bal} and D_{synaug} are fused to create a new dataset 'D_{real_syn}'. This blended dataset ensures fair learning across all classes and accurately models the distributions for classifier training. The Table 2 highlights all the symbols and notations used in this paper.

The synthetic data, replacing real data, must exhibit high quality, diversity, impartiality, and freedom from noisy samples. The challenge in synthetic data generation techniques lies in achieving data with such characteristics. It's crucial to note that synthetic data is not a silver bullet; its applicability is specific to the data, the model, and the domain.

3.1 Data balancing approach

In order to ensure that the minority class labels are equally represented, minority class labels are amplified and balanced with the help of many oversampling techniques. One such primitive technique is called ROS (random oversampling). ROS achieves class distribution balance by replicating existing minority class samples. ROS is not used in practical cases as the data generated is highly noisy and low in quality. ROS is a blind and uninformed sampling approach which simply duplicates the minority samples and does not increase the diversity and distribution of data samples. The authors in research work [24] pioneered the implementation of the informed, interpolation-based oversampling method known as SMOTE, which has proven effective even in the age of deep learning. The core principle of SMOTE involves augmenting the minority class by generating synthetic samples along the line segments connecting any (or all) of the k nearest minority neighboring data points.

SMOTE extends random oversampling by generating synthetic minority class samples, denoted as x_{new}. These are generated by choosing a random minority case, x_i, and creating a linear combination with a neighboring minority case, x_j, randomly selected from the k-nearest neighbors. This is done along a random point on the line connecting both cases, resulting in a new value $x_{new} = x_i + \in * (x_j - x_i)$, where $\in \sim [0,1]$. The hyper parameter k neighbors determine the number of neighboring cases. The other variants of SMOTE like adaptive synthetic sampling (ADASYN) also exist in the literature but the original SMOTE is the kind of benchmark oversampling technique as the other deviations from SMOTE doesn't necessarily beat their parent technique [25, 42, 55].Therefore, we have used SMOTE technique in this study for balancing the majority/minority distribution of dataset to achieve parity with respect to both the target classes. The dataset, balanced according to the target label after applying SMOTE, is illustrated in Eq. 1.

$$D_{bal} = SMOTE\left(D_{real}\right) \tag{1}$$

3.2 Data augmentation approach

Once the process of data balancing on target class is achieved, the now balanced, yet smaller dataset is passed through data augmentation funnel. The objective of data augmentation is to amplify and inflate the size of the smaller dataset so that its usefulness and application in machine learning downstream tasks for prediction, classification and decision making is bolstered. We have employed the services of deep generative modeling based generative adversarial networks to help improve the quality, size and representation of smaller datasets. The balanced dataset D_{bal} obtained in Eq. 1 is then fed into a generative model framework to produce a new synthetic augmented dataset D_{synaug} which is produced by applying CTGAN to D_{bal} from Eq. 1, as illustrated in Eq. 2.

$$D_{symaug} = CTGAN\left(D_{bal}\right) \tag{2}$$

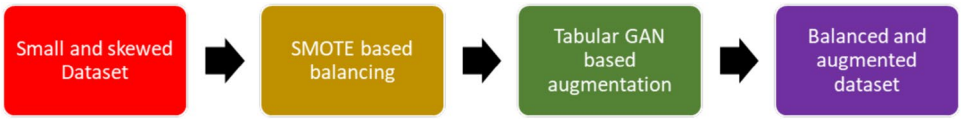

Fig. 1 Pipeline of hybrid synthetic data generation

In the context of conditional synthetic data augmentation, the objective is to produce a high-quality, debiased, useful, and augmented with diverse samples dataset. Therefore, we fuse the balanced datasets and synthetic dataset together in order to yield a new dataset $D_{real\ syn}$ such that output target labels $|Y_M|$ and $|Y_m|$ are comparable (i.e., IR ~ 1). This implies the dataset obtained in Eq. 2 which is an artificial version of the balanced dataset is fused with Eq. 1 to yield the final blended dataset which can be represented with the help of Eq. 3

$$D_{real\_syn} = D_{bal} + D_{synaug} \tag{3}$$

$D_{real\ syn}$ is the final and ready to be used new dataset on which a classifier C (such as a logistic regression, random forest, a knn, a decision tree, or a neural network) learns all classes fairly during training process of the classification algorithms and the distributions are accurately modeled (Fig. 1).

In the landscape of tabular data, creating synthetic data is not a trivial operation. Tabular datasets are challenging, in a sense, that, such datasets contain a mixture of continuous and discrete columns, also called as, numerical and categorical columns. Each of the type has its own intricacies and are taken care of accordingly. Numerical columns contain multi modal distributions which are usually non-gaussian, whereas, categorical columns are often imbalanced in which the major category overwhelms the minority rows. This creates severe mode collapse. It is difficult for the discriminator to notice minute variations in the data distribution that result from missing a minor category. For lesser classes, imbalanced data also results in inadequate training chances [51].

The pictorial representation in the feature space after data balancing and data augmentation of smaller and skewed dataset is shown with the help of Fig. 2.

3.3 Our presented hybrid approach: HDBA-MTD based shallow to deep intelligent oversampling

The presented HDBA-MTD approach consists of balancing module (SMOTE module) which is there to even majority/minority discrepancy with the help of generating synthetic minority samples. The balancing approach is followed by conditional tabular adversarial network that aims to mimic the probability distribution of data samples to yield completely new synthetic dataset. The generative modeling framework includes the generator and discriminator modules. The generator network takes random noise vector as input and produces a generated output. The generated samples are subjected to discriminator module which is designed

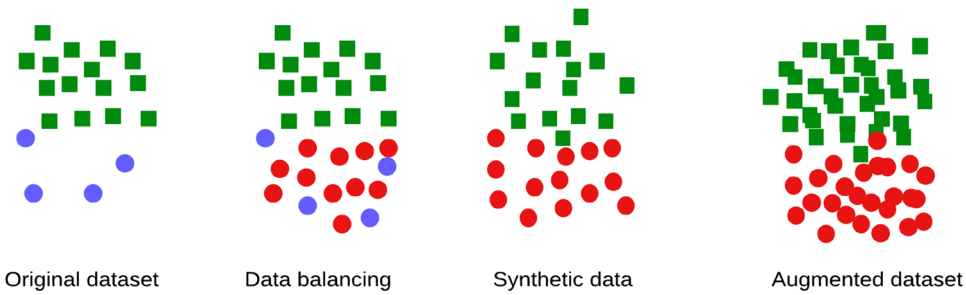

Fig. 2 Feature space of data samples by employing synthetic data generation process

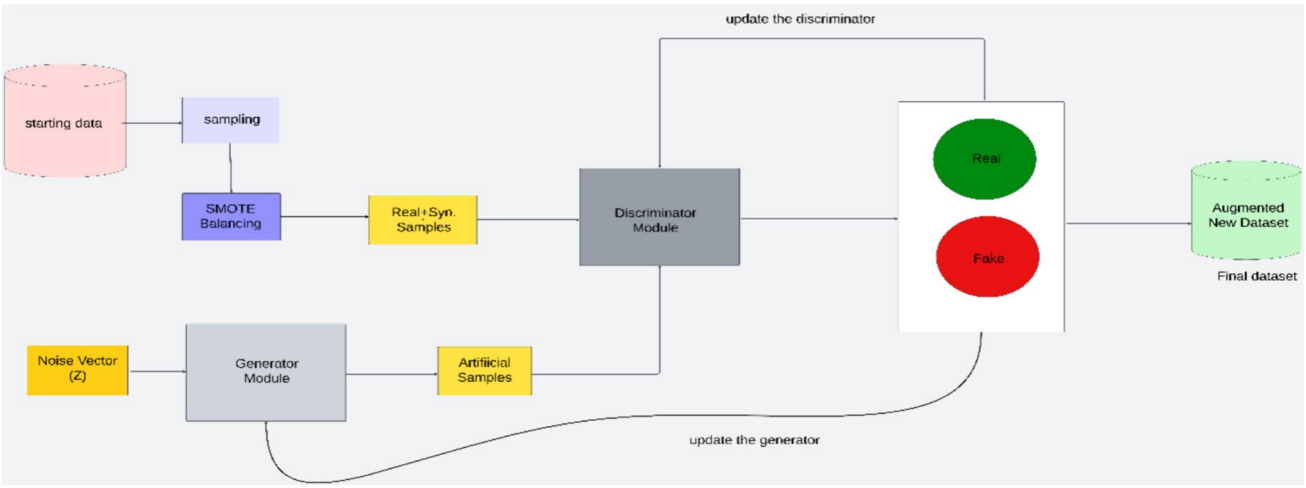

Fig. 3 Block diagram of HDBA-MTD approach

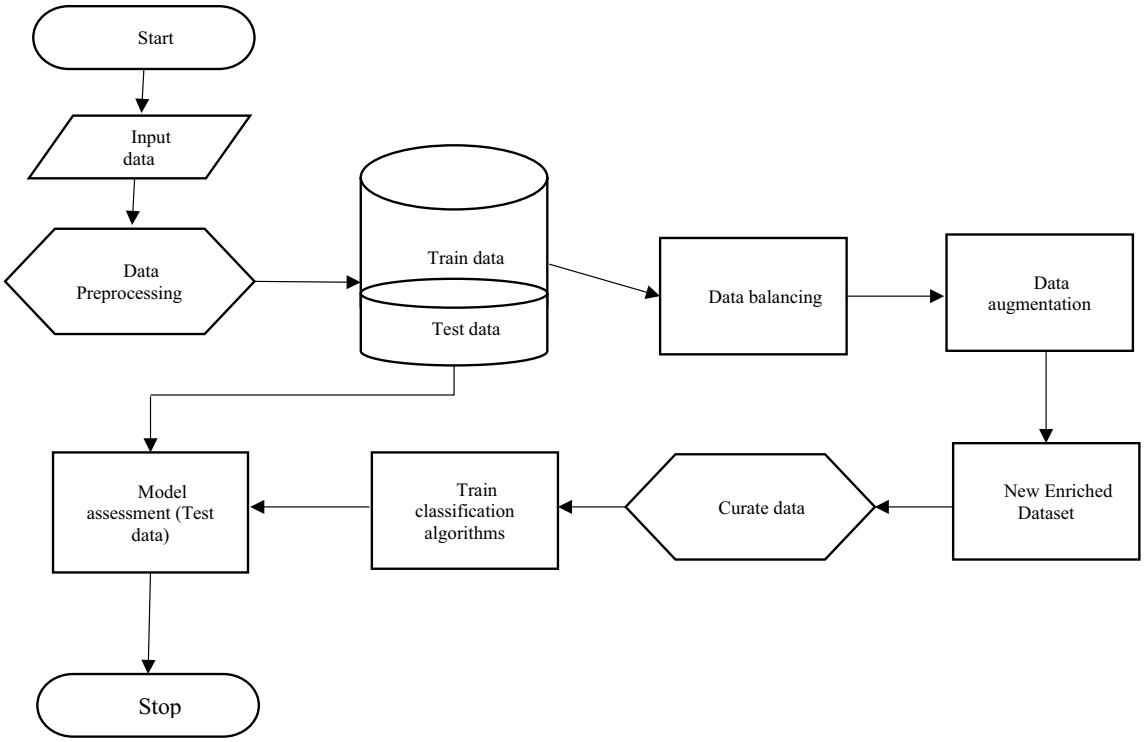

Fig. 4 Flowchart of HDBA-MTD approach for synthetic data generation

to differentiate between real data samples and fake data samples. The discriminator module takes two inputs. One is coming from the balancing SMOTE module which are treated as real samples and the other input is fed from the output of the generator module. The job of the discriminator is to categorise samples as fake and real based on the information it is receiving in the form of inputs. With many iterators and epochs both the modules of GAN architecture are updated in order to compete against each other in an adversarial fashion. The block diagram of the proposed approach is illustrated with the help of Fig. 3.

The aim is to enhance the generalization and performance of classification tasks by offering level playing field for both majority and minority classes, thereby alleviating potential issues of misleading accuracy arising from label imbalance in data-constrained settings. Illustrated with the help of flowchart in Fig. 4, our proposed framework based on the HDBA-MTD model comprises several interconnected phases designed to achieve these goals.

In the initial phase, we consider a tabular dataset with mixed data types and severe class imbalance. Before training any model, we preprocess the data to clean it, addressing issues such as missing values and inappropriate formats to ensure effective learning. Next, we employ a hybrid approach utilizing SMOTE to balance the target classes in the dataset, addressing the skewness commonly found in real-world data. The balanced dataset is then further processed using a conditional tabular GAN to generate additional data samples. These samples undergo fidelity checks to retain only high-quality data points. Subsequently, the synthetic dataset obtained from the GAN is merged with the balanced dataset to create a new, enriched, and diverse dataset. This dataset is then prepared for integration into the machine learning pipeline for classifier building. Before building the classifiers, the dataset undergoes further processing to ensure compatibility with machine learning algorithms, including normalization, standardization, and categorical data encoding. Finally, various machine learning classifiers are trained on the new enriched dataset, including support vector machine, random forest, decision trees, k-nearest neighbors etc. The performance of these classifiers is evaluated using benchmark datasets augmented with synthetic data to assess improvements in machine learning efficacy.

Our experimental findings surpass those of nearby studies by a significant margin, providing substantial evidence of the novelty of our work and justifies the introduction of balancing and augmentation modules for avoiding bias and increasing the diversity of data for generalisation of the model.

3.4 Benchmark methods

We take into consideration the baseline solution (no oversampling) of training a classifier on the severely imbalanced

real data in order to evaluate the proposed HDBA-MTD based oversampling approach. Additionally, we contrast with existing hybrid oversampling methods like SMOTE, synthetic minority oversampling generative adversarial network (SMOGAN), synthetic minority oversampling wasserstein generative adversarial network SMOwGAN. The imbalanced-learn package is utilised in the benchmark oversampling method implementations. The original SMOTE is the kind of benchmark oversampling technique as the other deviations from SMOTE doesn't necessarily beat their parent technique [25, 39, 42].Therefore, we have used SMOTE technique in this work for drawing comparisons and the existing related studies have also used SMOTE as a baseline approach.

3.5 Model agnostic metrics

In our work, we have focused on selecting model agnostic metrics which can be used for any classifier to evaluate the performance. We evaluate the performance of different classifiers like decision tree classifier (DTC), random forest classifier (RFC), k nearest neighbors (k-NN) and support vector machines (SVM) with respect to both classes of a binary classification problem using three metrics: accuracy, f1-score and balanced accuracy. The classification algorithms are executed 10 times, and the mean average of the evaluation metrics is documented in this research work. Accuracy is a misleading metric in the context of class imbalanced learning. Rest all metrics used here are suitable for class imbalanced datasets. The metrics are robust towards class imbalance and do not requires any different settings. Balanced accuracy focuses on TPR (true positive rate) and TNR (true negative rate) and is typically used in the context of imbalanced data learning in case of classification problem. Precision and Recall are the typical indicators that are used to reduce type I and type II errors of a confusion matrix in case of classification problem. Recall for the minority class is the metric we would like to optimize (e.g. in case of disease diagnosis, fraud detection etc.). The aim is to reduce false negatives (FN), thereby, maximising the true positive rate (TPR). For precision, in the context of spam classification, the aim should be to minimise the false positive rate (FPR), so, would like to reduce false positives (FP) in confusion matrix in order to maximise the precision metric. However, if both false positives and false negatives are equally significant in imbalanced datasets then both recall, and precision is used and f-measure (f1-score) is evaluated. It is noted that, larger values are preferable for accuracy, balanced accuracy, and f1 score.

Table 3 performance metrics

Metric	Formula	Ideal value
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	1
Recall/TPR/Sensitivity	$\frac{TP}{TP+FN}$	1
Precision	$\frac{TP}{TP+FP}$	1
Specificity/TNR	$\frac{TN}{TN+FP}$	1
Balanced Accuracy	$\frac{\text{Sensitivity+ Specificity}}{2}$	1
F-score (f1)	$\frac{2*Precsion*Recall}{Precision+Recall}$	1

3.6 Datasets and data organization

To assess the effectiveness of the proposed HDBA-MTD, comprehensive experiments were conducted on many real-life benchmark datasets. These datasets are publicly available on UCI ML and Kaggle repositories and are widely utilized for evaluating the performance of machine learning models in class imbalanced settings. Detailed information about these datasets is presented in Table 3.

3.7 Implementation

In classifier comparisons, choosing how to handle hyper parameters is crucial. Both individual classifiers and oversampling techniques have hyper parameters. The latter are comparatively more significant in this oversampling-focused investigation. As the proposed work is based on the hypothesis that synthetic data balancing and data augmentation can enhance the performance in data scarcity conditions. Therefore, for this reason, we refrain from tuning classification algorithms.

In our experiment, we divided the original datasets into 80%-20% train-test splits. The 20% test data remained untouched and was solely utilized for model evaluation. In addition of using the same SVM classifier for comparison with existing works like [23], we also employed many other classifiers to demonstrate the effectiveness of employing the proposed HDBA-MTD approach. This is detailed in the results and discussion section. As for the 80% train data, we partitioned the dataset based on majority and minority class labels. Given that most of the smaller datasets presented in Table 3 are balanced, we opted to reduce the minority class samples to a subset of 4%. The objective was to deliberately induce severe imbalance within the minority class labels, which were less than 5%, thereby creating a notably high imbalance ratio (IR). IR represents the ratio of the number of samples in the majority class to that in the minority class; the higher the IR, i.e., $IR > 20$, the more severe the imbalance [22]. Subsequently, this new 4% minority subset was subjected to balancing by SMOTE initially, followed by augmentation of the training dataset to enhance and enrich the overall balanced dataset, thereby facilitating effective classification performance. This utilization of the 4% minority rule draws inspiration from [23]. Given the nature of this

study, we maintain a consistent set of reasonable default hyper parameter settings across all experiments.

In general, the training of all methods is carried out on a single server with Intel(R) Core (TM) i9-12,900 clocked at a base speed of 2.40 GHz, 32 GB of memory and a graphics processing unit (GPU) with AMD Radeon RX 6500 M. The imbalanced-learn package was used in the trials to implement the benchmark oversampling techniques. We employed an open-source conditional tabular GAN from synthetic data vault (sdv) library [61] tailored for tabular datasets implementation and we fused the architecture as per the presented approach, making thoughtful modifications to ensure the creation of a high-quality and diverse dataset.

4 Results

This research paper delves into a comprehensive exploration of the fundamental limitations encountered in downstream tasks pertaining to machine learning and deep learning due to class imbalance and limited data availability. We introduce a novel hybrid technique to effectively address and overcome these challenges. Our focus is on exploring key research questions that illuminate potential improvements and advancements in addressing dilemmas related to data scarcity and data biasing on labels through synthetic data generation. This work demonstrates superior performance compared to the methods evaluated in [40], including SMOTE, SMOGAN, and SMOwGAN. The authors in [40] conducted a comparative analysis of these methods. Our approach outperforms all techniques discussed in their research. This section details the comparison, highlighting our method's enhanced performance over competing approaches.

RQ1: Can the proposed approach effectively improve classification algorithm performance in scenarios where the prediction output is adversely affected by majority class samples?

Algorithms learn from data, and when the data is biased in the favor of majority class, this bias is reflected in the prediction output. The presented approach employs an intelligent hybrid oversampling scheme that allows us to control the type and amount of synthetic data generated. We mitigate the influence of majority class samples by inflating the dataset with the desired number of data points from the minority feature space. Once the balancing of the samples is achieved, we augment the datasets with generative modeling technique for increasing the diversity of data to improve generalisation. As previously mentioned, machine learning algorithms only learn from what they encounter during training. Now, since we have increased the representation of data samples in the dataset, it is natural to expect improved classifier performance with respect to evaluation metrics. This improvement is highlighted in the results Table 5. Furthermore, supported by the illustration depicted in Fig. 5, we have substantiated the assertion that our hybrid oversampling HDBA-MTD approach effectively and efficiently enhances classification performance in scenarios with limited data availability.

RQ2: Can the proposed approach yield performance improvements over existing hybrid-based techniques when applied to diverse datasets?

To address RQ2, we opted for the widely utilized data balancing technique, SMOTE, as employed in the relevant studies, and designated no oversampling as the baseline. The selection of no oversampling as the baseline method is based on the rationale that if data augmentation fails to outperform this basic approach, it suggests that oversampling existing

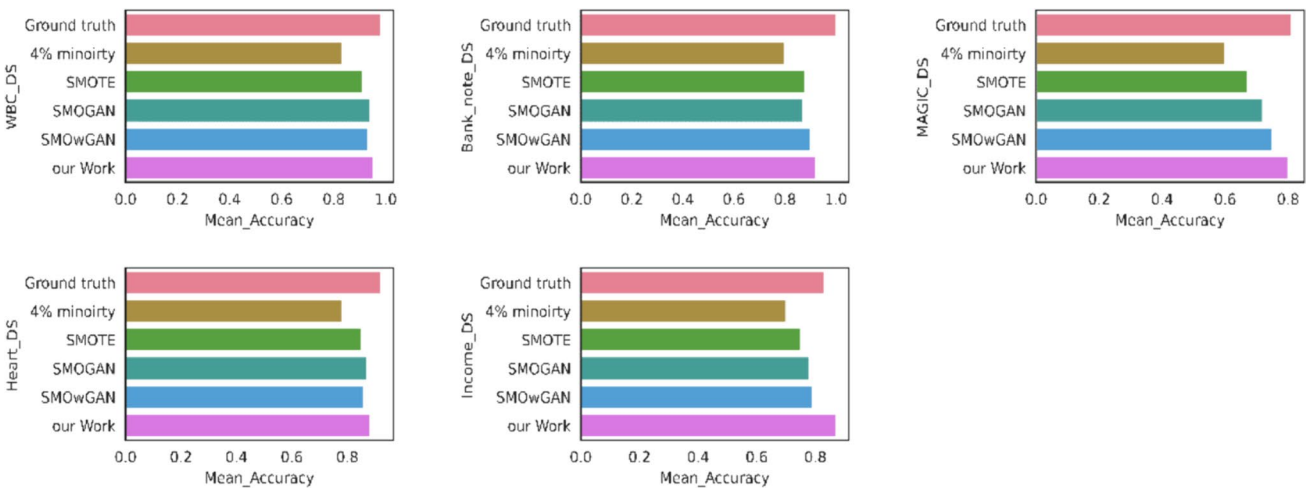

Fig. 5 Illustration of mean accuracy based on various datasets. The plots highlight the superior performance of our approach when data is severely imbalanced and less representative for efficient model training

Table 4 datasets used and their description

Dataset	#rows	#attributes	#Majority	#Minority	4% Minority	Artificial IR
WBC [56]	569	30	357	212	9	39
Bank Note [57]	1372	4	762	610	25	55
MAGIC [58]	19,020	10	12,332	6688	268	46
Heart Health [59]	1025	14	526	499	20	26
Adult Income [60]	48,842	15	37,155	11,687	468	79

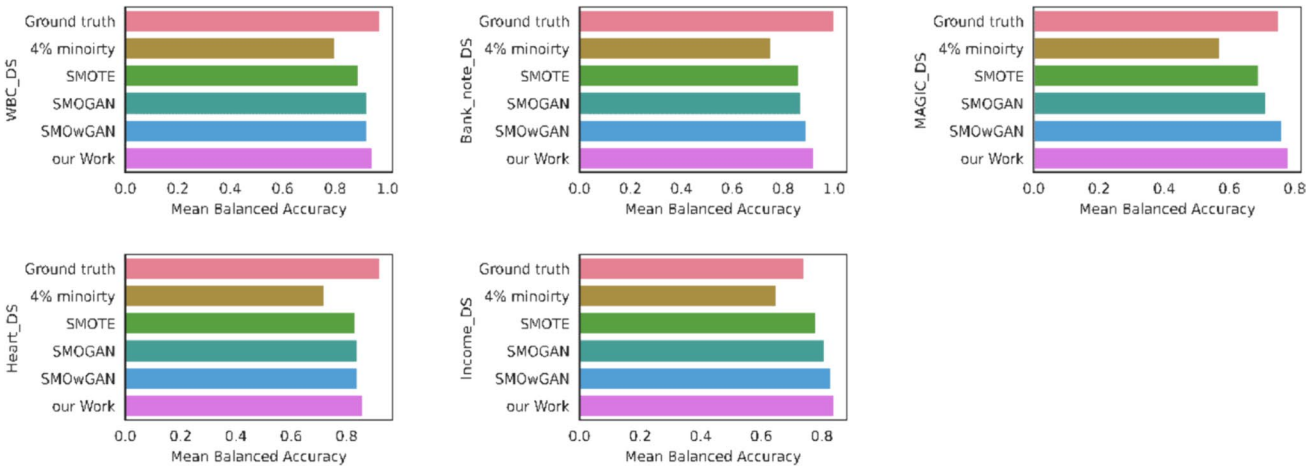

Fig. 6 comparison of mean balanced accuracy based on various datasets

Table 5 comparison of mean classification accuracy based on various datasets

Dataset	No oversampling	4% minority	SMOTE	SMOGAN	SMOwGAN	Our Work
Banknote	1.0	0.80	0.88	0.87	0.90	0.92
WBC	0.98	0.83	0.91	0.94	0.93	0.95
Magic	0.81	0.60	0.67	0.72	0.75	0.80
Heart	0.92	0.78	0.85	0.87	0.86	0.88
Income	0.83	0.70	0.75	0.78	0.79	0.81

data may not be necessary. Table 5 contains comparative results for many datasets. However, as can be observed, our presented approach beats the other methods and validates that it does comparably well in such settings.

RQ3: Does this approach perform effectively when dealing with both categorical and numerical columns?

The Table 4 presents the datasets utilized in this study, emphasizing benchmark datasets that feature a combination of numerical and categorical columns. Our approach employs the use of mode specific normalization for handling multi modal data while existing techniques in the literature use min–max normalisation to model the numerical columns and hence fail to take care of multiple modes. This indicates that current techniques are unable to effectively represent the various data patterns within numerical columns. In addition to this, the CTGAN handles categorical columns separately and therefore demonstrates superior capability in handling numerical and categorical columns compared to existing approaches.

RQ 4: Does the presented approach help to improve the precision, recall, TPR and TNR?

When it comes to dealing with class imbalance and data scarcity circumstances, improving on TPR, TNR, precision, recall collectively realised in the form of f1 score are significant indicators of true model performance. When both false positive and false negatives are crucial, f1 score is considered as it encompasses individual metrics like precision and recall together. Therefore, in addition to balanced accuracy, we have also adopted f1 score in this work. Our approach has done reasonably well in comparison to existing techniques and the improvement is shown with the help of Fig. 6 and

Table 6 comparison of mean f1 score based on various datasets

Dataset	No over-sampling	4% minority	SMOTE	SMOGAN	SMOwGAN	Our Work
Banknote	0.96	0.85	0.88	0.87	0.88	**0.91**
WBC	0.96	0.81	0.87	0.88	0.86	**0.92**
Magic	0.65	0.59	0.68	0.68	0.69	**0.72**
Heart	0.89	0.81	0.85	0.82	0.83	**0.85**
Adult	0.76	0.70	0.78	0.81	0.83	**0.85**

The bold text indicates that our approach outperforms the other methods, demonstrating superior performance

Table 6. The details of the improvements is discussed in the next section.

RQ5: Does the presented hybrid approach outperform all the other techniques under different experimental settings?

The presented method has majority of the times shown superior performance compared to other techniques across all datasets and experimental conditions, as illustrated in Tables 5, 6 and respective Figs. 5 and 6. However, our observations indicate that in datasets where the number of majority and minority samples is balanced, natural data often achieves comparable or better performance than oversampling techniques alone. This suggests that while artificial data can enhance existing datasets, it cannot fully substitute real data, at least not at present. Additionally, our findings reveal that oversampling methods have effectively improved classification performance in scenarios where data instances are imbalanced and limited in quantity.

5 Discussion

Data scarcity and class imbalance present formidable obstacles in machine learning for classification and prediction tasks. Limited data availability hinders the effectiveness of algorithms, affecting insights across research domains. While balanced datasets are optimal, they are rare; imbalance is common, skewing towards dominant groups. Synthetic data is used to address these issues. Traditional evaluation methods struggle with unbalanced data, as algorithms tend to favor majority classes, leading to misleading metrics [11]. This imbalance challenges accurate classification of underrepresented classes, often resulting in the accuracy paradox.

Algorithms derive their knowledge from data, and when the data is skewed towards the majority class, this bias is evident in the prediction results. Our approach utilizes an intelligent hybrid oversampling method, enabling precise control over the type and quantity of synthetic data produced. By incorporating an appropriate number of data points from the minority feature space, we alleviate the impact of majority class samples on the dataset. After achieving sample balance, we enhance the datasets using generative modeling techniques to promote diversity and enhance generalization. The enhancement is observed and is presented as answered to various RQs.

In this research, we into consideration the baseline solution (no oversampling) of training a classifier on the severely imbalanced real data in order to evaluate the proposed HDBA-MTD based oversampling approach. Choosing no oversampling as the baseline method is grounded in the idea that if data augmentation does not surpass this basic approach, it implies that oversampling existing data might be unnecessary and is not improving the performance of downstream machine learning tasks. Additionally, we contrast with existing hybrid oversampling methods like SMOTE, SMOGAN, SMOwGAN in order to test the superiority of our approach.

This work focuses on selecting model-agnostic metrics for evaluating classifier performance. We use accuracy, f1-score, and balanced accuracy to assess binary classification models. Accuracy may be misleading for imbalanced data, while other metrics remain robust. Balanced accuracy considers true positive rate (TPR) and true negative rate (TNR) and is suitable for imbalanced data. Precision and Recall help manage type I and type II errors, with Recall prioritizing the minority class. Precision aims to minimize false positive rate (FPR), crucial in contexts like spam detection. For balanced importance of both false positive (FP) and false negative (FN), f-measure (f1-score) is used. In our research, we assess our proposed approach using multiple metrics, including accuracy, balanced accuracy, and F1-score. This comprehensive evaluation aims to prevent bias toward any single metric, demonstrating our commitment to equally prioritizing true positive rate, true negative rate, while minimizing false positive rate, false negative rate, as well as type I and type II errors. Figures 5, 6, and Table 6 present the results for accuracy, balanced accuracy, and F1-score, respectively. These metrics encompass all the derivatives of the confusion matrix (TP, TN, FP, and FN). These results underscore the superior performance of our approach compared to existing methods.

In this work, we have seen that our approach is outperforming existing synthetic data generation techniques on all performance evaluation metrics. This is attributed to the fact

that tabular data contains non-Gaussian and multimodal continuous columns and highly skewed discrete columns. In this research we have fused SMOTE with CTGAN architecture because CTGAN employs mode-specific normalization to address distributions that are non-Gaussian and multimodal in nature. Moreover, it utilizes a conditional generator and training-by-sampling to tackle imbalanced discrete columns. This helps us to model the nuances and heterogeneity of structured data attributes better as compared to the existing GAN architecture used in the related studies.

Choosing hyper parameters is crucial in classifier comparisons, especially in oversampling-focused investigations where oversampling techniques have significant hyper parameters. Our work hypothesizes that synthetic data balancing and augmentation enhance performance in data scarcity, hence we avoid tuning classification algorithms. In our experiment, we split datasets into 80%-20% train-test splits, reserving the 20% for evaluation. We employ various classifiers besides SVM to demonstrate the effectiveness of our approach. We deliberately induce severe imbalance in minority class labels, reducing them to less than 5% to create a high imbalance ratio. In our research, we have artificially created a very severe majority-minority divide, precisely to 4%. This means we have retained only 4% of minority samples in relation to the majority data samples from the 80% training data. The 20% test data is untouched and is used solely for model evaluation. The 4% minority data samples are utilized in the study [40] against which we are comparing our approach. For the sake of maintaining consistent parameters for evaluation, we have employed the same percentage. However, this also ensures that we have a high imbalance ratio (IR > 20) in the dataset, allowing for the observation of the negative effects of data imbalance on classification performance. We then balance the minority subset using SMOTE and augment the training dataset for improved classification. We maintain consistent default hyper parameter settings across all experiments. In our research, we have not altered the hyper parameter settings of either the classifiers or the CTGAN architecture. This decision was made to avoid biasing the evaluation metrics. Tuning hyper parameters often improves model performance. This makes it challenging to pinpoint the source of the performance improvement. Whether it stems from our presented hybrid approach or from the influence of the hyper parameters acting as proxies and affecting model outcomes.

In this research, we have used many tree-based classifiers like DTC, RFC, k-NN apart from the SVM. The SVM was used by the existing studies and we have selected above mentioned tree-based classifiers as well. The inclusion of DTC and k-NN is motivated by their expected sensitivity to the quality of generated minority class instances. DTCs are prone to greedy growth and overfitting, particularly given that the only synthetic samples in the training data belong to the minority class. This makes them susceptible to fitting bogus patterns in the generated data, potentially resulting in poor performance on the test set. Successful oversampling techniques in conjunction with a DTC, therefore, indicate high-quality crafted minority class instances. Likewise, the effectiveness of k-NN in oversampling relies on how closely the generated minority samples match the true minority samples in the test set. This is an implicit indicator of how well the oversampling technique approximates the distribution of minority class instances. The improvement in different performance evaluation metrics with respect to tree-based classifiers like RFC, DTC and k-NN as illustrated in the Fig. 7 validate that the presented data oversampling approach is performing satisfactory as such algorithms are sensitive to the quality of the synthetic data generated.

When evaluating the utility and fidelity of synthetic data generated by generative models, including the approach

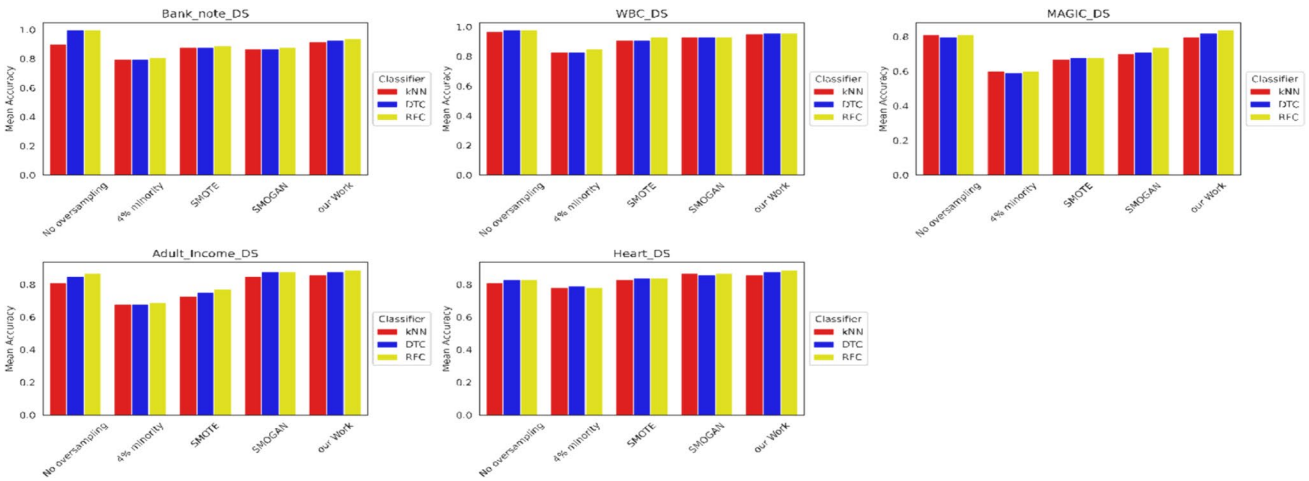

Fig. 7 comparison of mean accuracy of various tree-based classifiers (k-NN, DTC, RFC) based on different datasets

utilized in this research, a critical challenge emerges regarding whether the synthetic data truly adds value. The efficacy of machine learning, particularly in the realm of deep generative models, is assessed through the train synthetic and test real (TSTR) paradigm. Under this paradigm, classification algorithms are not directly trained on real data but rather on synthetic-enriched datasets. If the data generated by generative models fails to enhance dataset diversity and instead comprises redundant and noisy instances, this can adversely impact model evaluation metrics when assessed on real test data. However, the robustness of the presented HDBA-MTD architecture enables effective handling of sampling under various conditions. Our results validate the usefulness of the generated data, as they demonstrate favorable outcomes despite evaluation being conducted within the TSTR framework.

The results obtained from accuracy, balanced accuracy, and f1 scores metrics underscore the significance of representative data, defined as the minimum number of samples necessary for effective learning. In datasets such as WBC, Heart_health, and bank_note, where the majority and minority instances are comparable and nearly balanced, the original data performs well and is not surpassed by synthetic data, as evidenced in the corresponding figures. However, upon inducing an artificial majority-minority divide, our approach consistently outperforms techniques employed in existing studies. Conversely, in datasets like MAGIC and Income, which are naturally imbalanced, the impact on performance metrics—accuracy, balanced accuracy, and F1 score—is notable. While the use of synthetic data generation approaches improves performance compared to baseline ground truth, our approach still outshines competitors due to architectural enhancements in the CTGAN framework. This suggests that in datasets with sufficient balance and diversity for effective learning, synthetic data generation may be unnecessary. Furthermore, this highlights the limitations of deep generative models in efficiently modeling real data with scarce datasets. Our study demonstrates that the CTGAN architecture faces challenges in modeling smaller datasets (e.g., WBC, bank_note, and Heart_health) when compared with the baseline (ground truth), while larger datasets (e.g., MAGIC, Income) are better suited for real data modelling and can be observed in result Tables 5 and 6. Exploring the optimal amount of starting real data for GAN-based architectures presents a promising research avenue for enhancing effective learning and efficient modeling of real data.

6 Conclusion and future research work

This paper investigates the significance of synthetic data generation in data scarcity dilemmas. We validate by empirical investigation that the phenomenon of class imbalance and limited nature of data severely hampers the performance of classification tasks in real-world datasets and representation for effective learning. To overcome these barriers, we introduced a novel hybrid oversampling technique based on smote and tabular GAN called CTGAN architecture and conducted comparisons with established oversampling techniques, along with a baseline scenario of no oversampling. The empirical analysis encompassed several benchmark datasets, employing different classification algorithms and evaluating performance using model agnostic classification metrics. Our findings indicate that HDBA-MTD based hybrid oversampling exhibits very favorable results when compared to other oversampling methods, surpassing no oversampling, SMOTE and vanilla GAN based SMOTE-style methods. The proposed approach ensures level playing field for all groups through synthetic data generation scheme based on conditional sampling of minority groups. The SMOTE based on interpolation mechanism acts as a data balancing approach and helps to synthesise the local information. In doing so, we addressed class imbalance while avoiding potential bias and accordingly modeled the synthetic data points. The balancing approach is followed by data augmentation for considering the global information and estimate the whole probability distribution of the balanced dataset. This generative modeling-based GAN augmentation ensures increasing the representation and enriching the data samples for effective learning. We have observed that when the dataset is comparably balanced and there are enough samples for representation, oversampling approaches have not significantly outperformed the baseline. However, when either or both conditions are not in conjunction, oversampling methods thrive. In this work, we have artificially induced minority in data samples to test the oversampling performance and we observed among the existing approaches, our presented approach is performing better in majority of the circumstances.

As already discussed in previous sections synthetic data is meant to tackle the data scarcity problems provided with less data or imbalanced data but it faces challenges in various aspects like the model architecture, data fidelity, and complex nature of tabular data columns etc. Key research directions for further evolution of synthetic data synthesis for modeling tabular data are succinctly summarized below.

- In this work, we emphasize that generating synthetic data for tabular datasets is not a straightforward task. This suggests that each tabular dataset possesses unique characteristics, with variations existing not only between datasets but also within individual columns. Consequently, generalizing a specific data synthesis technique becomes challenging, continuing to be an area of active interest among researchers in the field. This contrasts

with image data, which tends to be more homogeneous in nature.

- This research underscores the applicability of synthetic data modeling to various classification problems plagued by data scarcity issues. Synthetic data can play a crucial role in numerous application areas, including disease detection, anomaly detection, fraud detection, intrusion detection, and failure prediction, among others. By supplementing real data and augmenting datasets, synthetic data enhances data diversity, thereby improving downstream classification tasks. Future work in this direction could focus on employing this hybrid model for dedicated application areas.
- Our research in this paper focuses solely on binary classification problems. Future studies in this domain, particularly concerning mixed data types within tabular data, can extend to multi-class and multi-label settings. Such investigations would also address bias issues associated with class imbalance and data scarcity challenges.
- Given the constraints imposed by data regulations and the sensitive nature of data, synthetic data modeling emerges as a vital tool for safeguarding data privacy. Research in this domain holds considerable significance for applications dealing with sensitive data.
- The effectiveness and fidelity of the generated artificial data directly influence the learning process, potentially resulting in subpar performance of models in downstream machine learning tasks. Investigating this aspect further constitutes a significant research direction. It is imperative that the synthetic data generated maintains both syntactic and semantic integrity.

Funding None.

Data availability Data is made available on request.

Declarations

Conflict of interest The authors declare no conflict of interest.

Informed consent Not applicable.

References

1. Vatansever S et al (2021) Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions. Med Res Rev. https://doi.org/10.1002/med.21764
2. Kaisler S, Armour F, Espinosa JA, Money W (2013) Big data: Issues and challenges moving forward. Proc the Annu Hawaii Int Conf Syst Sci. https://doi.org/10.1109/HICSS.2013.645
3. Pandey R, Gupta A, Pandey A (2022) The internet of medical things (IoMT) and telemedicine frameworks and applications. https://doi.org/10.4018/978-1-6684-3533-5.
4. Siddiqui T, Mustaqeem M (2023) Performance evaluation of software defect prediction with NASA dataset using machine learning techniques. Int J Info Technol (Singapore). https://doi.org/10.1007/s41870-023-01528-9
5. Lv G et al (2023) Laser ultrasonics and machine learning for automatic defect detection in metallic components. NDT E Int. https://doi.org/10.1016/j.ndteint.2022.102752
6. Afriyie JK et al (2023) A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decis Anal J. https://doi.org/10.1016/j.dajour.2023.100163
7. Zhai YJ, Zhang Y, Liu HZ, Zhang ZR (2023) Multi-angle support vector survival analysis with neural tangent kernel study. Arab J Sci Eng. https://doi.org/10.1007/s13369-022-07540-8
8. Salim A, Juliandry LR, Moniaga JV (2022) General pattern recognition using machine learning in the cloud. Proced Comput Sci. https://doi.org/10.1016/j.procs.2022.12.170
9. Jiao R, Li C, Xun G, Zhang T, Gupta BB, Yan G (2023) A context-aware multi-event identification method for nonintrusive load monitoring. IEEE Trans Consum Electron. https://doi.org/10.1109/TCE.2023.3236452
10. Ganganwar V, Rajalakshmi R (2023) Enhanced Hindi aspect-based sentiment analysis using class balancing approach. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-023-01430-4
11. Bhat KA, Sofi SA (2024) Class imbalanced problem: Taxonomy, open challenges, applications and state-of-the-art solutions. China Commun. https://doi.org/10.23919/JCC.EA.2022-0448.202401
12. Vega-Márquez B, Rubio-Escudero C, Riquelme JC, Nepomuceno-Chamorro I (2020) Creation of synthetic data with conditional generative adversarial networks. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-030-20055-8_22
13. Hasanin T, Khoshgoftaar TM, Leevy JL, Bauder RA (2019) Severely imbalanced big data challenges: investigating data sampling approaches. J Big Data. https://doi.org/10.1186/s40537-019-0274-4
14. Mumuni A, Mumuni F (2022) Data augmentation: a comprehensive survey of modern approaches. Array. https://doi.org/10.1016/j.array.2022.100258
15. Fonseca J, Bacao F (2023) Tabular and latent space synthetic data generation: a literature review. J Big Data. https://doi.org/10.1186/s40537-023-00792-7
16. Kaur P, Gosain A (2022) Issues and challenges of class imbalance problem in classification. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-018-0251-8
17. Khalifa NE, Loey M, Mirjalili S (2022) A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif Intell Rev. https://doi.org/10.1007/s10462-021-10066-4
18. Hasib KM et al (2020) A survey of methods for managing the classification and solution of data imbalance problem. J Comput Sci. https://doi.org/10.3844/JCSSP.2020.1546.1557
19. Das S (2024) A new technique for classification method with imbalanced training data. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-024-01740-1
20. Sampath V, Maurtua I, Aguilar Martín JJ, Gutierrez A (2021) A survey on generative adversarial networks for imbalance problems in computer vision tasks. J Big Data. https://doi.org/10.1186/s40537-021-00414-0
21. Sauber-Cole R, Khoshgoftaar TM (2022) The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. J Big Data. https://doi.org/10.1186/s40537-022-00648-6
22. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N (2018) A survey on addressing high-class imbalance in big data. J Big Data. https://doi.org/10.1186/s40537-018-0151-6

23. Mohammed R, Rawashdeh J, and Abdullah M (2020) Machine learning with oversampling and undersampling techniques: overview study and experimental results," 2020 11th International Conference on Information and Communication Systems, ICICS 2020, pp. 243–248, 2020, https://doi.org/10.1109/ICICS49469.2020.239556.
24. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/JAIR.953
25. Engelmann J, Lessmann S (2021) Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.114582
26. Rodriguez-Almeida AJ et al (2023) Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2022.3196697
27. Yang Z, Li Y, Zhou G (2023) TS-GAN: time-series GAN for sensor-based health data augmentation. ACM Trans Comput Healthc. https://doi.org/10.1145/3583593
28. Lu H, Du M, Qian K, He X, Wang K (2022) GAN-based data augmentation strategy for sensor anomaly detection in industrial robots. IEEE Sens J. https://doi.org/10.1109/JSEN.2021.3069452
29. Zhang Y et al (2023) GAN-based one dimensional medical data augmentation. Soft comput. https://doi.org/10.1007/s00500-023-08345-z
30. Fedoruk O, Klimaszewski K, Ogonowski A, and Możdżonek R (2024) "Performance of GAN-based augmentation for deep learning COVID-19 image classification," In: International workshop on machine learning and quantum computing applications in medicine and physics: wmlq2022. https://doi.org/10.1063/5.0203379.
31. Al Khalil Y, Amirrajab S, Lorenz C, Weese J, Pluim J, Breeuwer M (2023) Reducing segmentation failures in cardiac MRI via late feature fusion and GAN-based augmentation. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2023.106973
32. Bhattacharya D, Banerjee S, Bhattacharya S, Uma Shankar B, Mitra S (2020) GAN-based novel approach for data augmentation with improved disease classification. https://doi.org/10.1007/978-981-15-1100-4_11.
33. Bhat S, Hortal E (2021) GAN-based data augmentation for improving the classification of EEG signals. ACM Int Conf Proc Ser. https://doi.org/10.1145/3453892.3461338
34. Motamed S, Rogalla P, Khalvati F (2021) Data augmentation using generative adversarial networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. Inform Med Unlocked. https://doi.org/10.1016/j.imu.2021.100779
35. Haruna Y, Qin S, Mbyamm Kiki MJ (2023) An improved approach to detection of rice leaf disease with GAN-based data augmentation pipeline. Appl Sci (Switzerland). https://doi.org/10.3390/app13031346
36. S. Sundaram and N. Hulkund, "GAN-based Data Augmentation for Chest X-ray Classification," Jul. 2021, Accessed: Apr. 27, 2024. [Online]. Available: https://arxiv.org/abs/2107.02970v1
37. Kiyoiti FH, Tanaka S, Aranha C, Lee WS and Suzuki T (2019) Data augmentation using GANs," *Proc Mach Learn Res*, vol. XXX, pp. 1–16, Accessed 27 Apr 2024. [Online]. Available: https://arxiv.org/abs/1904.09135v1
38. Manousakas D, Serg S, and Aydöre S (2023) On the Usefulness of Synthetic Tabular Data Generation,"Accessed27 Apr 27 2024. [Online]. Available: https://arxiv.org/abs/2306.15636v1
39. Sharma A, Singh PK, Chandra R (2022) SMOTified-GAN for class imbalanced pattern classification problems. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3158977
40. Lan ZC, Huang GY, Li YP, Rho S, Vimal S, Chen BW (2023) Conquering insufficient/imbalanced data learning for the internet of medical things. Neural Comput Appl. https://doi.org/10.1007/s00521-022-06897-z
41. Scott M, Plested J (2019) GAN-SMOTE: a generative adversarial network approach to synthetic minority oversampling for one-hot encoded data *ICONIP2019 Proceedings*, vol. 15, no. 2
42. Dablain D, Krawczyk B, Chawla NV (2023) DeepSMOTE: fusing deep learning and smote for imbalanced data. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2021.3136503
43. Cheah PCY, Yang Y, Lee BG (2023) Enhancing Financial fraud detection through addressing class imbalance using hybrid SMOTE-GAN techniques. Int J Financ Stud. https://doi.org/10.3390/ijfs11030110
44. Xu L, Skoularidou M, Cuesta-Infante A, and Veeramachaneni K (2019) Modeling tabular data using conditional GAN. In: Advances in neural information processing systems
45. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F (2019) Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf Sci (N Y). https://doi.org/10.1016/j.ins.2017.12.030
46. Douzas G, Bacao F (2018) Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2017.09.030
47. Openai IG (2016) NIPS 2016 tutorial: generative adversarial networks," Accessed 27 Apr 2024. [Online]. Available: https://arxiv.org/abs/1701.00160v4
48. Xu L and Veeramachaneni K (2018) Synthesizing Tabular Data using Generative Adversarial Networks,", Accessed 28 May 2024. [Online]. Available: http://arxiv.org/abs/1811.11264
49. Cheon MJ, Lee DH, Park JW, Choi HJ, Lee JS, Lee O (2021) CTGAN VS TGAN? Which one is more suitable for generating synthetic EEG data. J Theor Appl Inf Technol 99(10):2359–2372
50. Baowaly MK, Lin CC, Liu CL, Chen KT (2019) Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocy142
51. Majeed A, Hwang SO (2023) CTGAN-MOS: Conditional generative adversarial network based minority-class-augmented oversampling scheme for imbalanced problems. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3303509
52. Singh PS, Singh VP, Pandey MK, Karthikeyan S (2022) Enhanced classification of hyperspectral images using improvised oversampling and undersampling techniques. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-021-00676-0
53. Kancharapu R, Ayyagari SN (2024) Suicidal ideation prediction based on social media posts using a GAN-infused deep learning framework with genetic optimization and word embedding fusion. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-023-01725-6
54. Iliyasu AS, Deng H (2022) N-GAN: a novel anomaly-based network intrusion detection with generative adversarial networks. Int J Inf Technol (Singapore). https://doi.org/10.1007/s41870-022-00910-3
55. Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2019.105662
56. "Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository." Accessed 28 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
57. "Banknote Authentication - UCI Machine Learning Repository." Accessed 28 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/267/banknote+authentication
58. "MAGIC Gamma Telescope - UCI Machine Learning Repository." Accessed 28 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope

59. “Statlog (Heart) - UCI Machine Learning Repository.” Accessed 28 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/145/statlog+heart
60. “Adult - UCI Machine Learning Repository.” Accessed 28 May 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/2/adult
61. “GitHub - sdv-dev/CTGAN: Conditional GAN for generating synthetic tabular data.” Accessed 27 Apr 2024. [Online]. Available: https://github.com/sdv-dev/CTGAN