Sauber-Cole and Khoshgoftaar
Journal of Big Data (2022) 9:98
https://doi.org/10.1186/s40537-022-00648-6

Journal of Big Data

SURVEY

Open Access

The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey

Rick Sauber-Cole* and Taghi M. Khoshgoftaar

*Correspondence:
rsaubercole2013@fau.edu

Florida Atlantic University, Boca Raton, FL 33431, USA

Abstract

The existence of class imbalance in a dataset can greatly bias the classifier towards majority classification. This discrepancy can pose a serious problem for deep learning models, which require copious and diverse amounts of data to learn patterns and output classifications. Traditionally, data-level and algorithm-level techniques have been instrumental in mitigating the adverse effect of class imbalance. With the recent development and proliferation of Generative Adversarial Networks (GANs), researchers across a variety of disciplines have adapted the architecture of GANs and implemented them on imbalanced datasets to generate instances of the underrepresented class(es). Though the bulk of research has been centered on the application of this methodology in computer vision tasks, GANs are likewise being appropriated for use in tabular data, or data consisting of rows and columns with traditional structured data types. In this survey paper, we assess the methodology and efficacy of these modifications on tabular datasets, across domains such network traffic classification and financial transactions over the past seven years. We examine what methodologies and experimental factors have resulted in the greatest machine learning efficacy, as well as the research works and frameworks which have proven most influential in the development of the application of GANs in tabular data settings. Specifically, we note the prevalence of the CGAN architecture, the optimality of novel methods with CNN learners and minority-class sensitive measures such as F1 score, the popularity of SMOTE as a baseline technique, and the improved performance in the year-over-year use of GANs in imbalanced tabular datasets.

Keywords: Class imbalance, Generative adversarial networks, Tabular data, Deep learning

Introduction

The proliferation of machine learning methods in many areas of industry and academia has acutely demonstrated the difficulty posed by class imbalance. Class imbalance occurs when the prevalence of one class or a handful of classes in a dataset outweighs the prevalence of another class (or classes) in the dataset. The class or classes which enjoy this prevalence are known as the majority classes, while the other class or classes which are relatively less frequent are known as the minority classes. In most settings, the minority

class is the class of interest to the practitioner. The primary effect of class imbalance is to bias the learner into the majority classification, to the detriment of identification of minority classes. The phenomenon has been studied extensively in the context of traditional machine learning [1–4], and many corrective methods have been proposed.

The problem is by no means made simpler by the rise of deep learning. With their black-box approach to modeling relationships amongst explanatory variables, and between explanatory variables and class labels or numeric outcomes, deep learning models have posed unique challenges for researchers and practitioners using imbalanced data. This is especially true because the method by which deep neural networks update their weights disproportionately favors the majority class [5], and because the extensive architecture of a deep neural network could lead to the memorization of randomly replicated instances. A steady stream of literature in the past decade has answered the call to action with diagnoses and proposed solutions [6–8].

Concurrently, Generative Adversarial Networks (GANs) proposed by Goodfellow et al. [9] have garnered a great deal of interest. Given a corpus of training data, these frameworks generate synthetic instances through the adversarial interplay of two or more networks consisting of Generators and Discriminators. Because GANs generate instances of the class that do not represent random replications of existing instances without removing valuable information in existing instances of data, they overcome the limitations of traditional techniques such as data sampling discussed in a later section. Moreover, many synthetic-based popular improvements on oversampling (such as SMOTE) use nearest neighbor linear interpolations to generate new instances. This localized approach can lead to meaningful improvements for lower-dimensional problems, but as Scott and Plested [10] note, because of its "feature space" rather than "data space" orientation, SMOTE is a suboptimal alternative for research problems in high-dimensional space.

In combination with the observations relating to class imbalance mentioned above, many researchers have investigated the feasibility of using GAN-derived architectures to generate synthetic minority instances to rebalance datasets for training. The goal of this paper is to offer an in-depth survey of those techniques, with a special focus on tabular datasets. Even though to date the prevalence of such methods has been relatively contained, the significant spread of GANs, deep learning, and big data architectures all but guarantee the prominence of generative methods as data augmenters and class imbalance rectifiers in a researcher's toolkit.

The remainder of this paper is structured as follows. We will first provide an operating definition for tabular data in our scope, and review the literature related to the topics of class imbalance and GANs. Subsequently, we will begin our review of the intersection of these two sets. Initially, we review the basic modifications to the GAN framework which have made it suitable for its application in correcting class imbalance. Furthermore, we look at the deployment of these methodologies in two subject domains: cybersecurity and financial transactions. Outside the umbrella of these two domains, we examine similar research conducted on other miscellaneous topics. Turning our attention away from the data-generation-to-learner-classification pipeline which characterizes the precedent sections, we then discuss a handful of research efforts that touch on the general elements in the GAN methodology which optimize performance. This includes the use

of Wasserstein loss to train the GAN network, a guide to choosing the correct evaluation metric with GANs given a variable level of class imbalance, and a systematic way to evaluate the quality of synthetic instances produced by GANs given the relative imbalance of different classes. Our exhaustive search for peer-reviewed research works related to the above topics concluded on September 23, 2021.

From here, we offer a meta-analysis of research findings and performance evaluation, and mine the citation network created by these research works for insights and patterns. With respect to this latter investigation, we will witness the prevalence of two architectures, CGAN and BAGAN (or CGANs with modified adversarial players) as the favored GAN methodologies, and the pronounced popularity of SMOTE as a baseline method. We will see experimental evidence of the ability of GANs to correct for class imbalance, as 83% of unique experiments among our researched papers boasted evaluation metrics for machine learning efficacy results where experimental methods outperformed baseline methods. Moreover, the degree of dominance between novel synthetic approaches and baseline approaches has been increasing year over year since the introduction of GANs into tabularized data. The advantage of novel methods will be most pronounced using F1 score and Ranking evaluation measures, and least noticeable with specificity and AUC. Likewise, the use of GANs for class imbalance in tabular environments has achieved greatest potency when classified with CNN and MLP learners and has proven least effective when paired with Naïve Bayesian and SVM learners. In our survey of the current landscape, we note, despite many significant advances in recent years, the numerous research opportunities available in the field, particularly as it relates to the application of GANs to tabular multi-class settings, and in determination of the optimal percent or level of GAN-generated instances.

Related concepts

Tabular data

Though the concept is well understood, given its centrality to the theme of this survey, we briefly examine the nature of and provide an operating definition to "tabular data". In general, tabular data consists of rows and columns, where the columns represent various attributes of a domain, and the rows represent instances of that domain. For example, in the realm of financial fraud detection, a row would represent a financial transaction, and the various columns would represent the attributes which characterize a transaction, such as time of day, amount in question, item purchased, location purchased, etc. These instances need not necessarily be unique in the final resultant dataset (that is, there could be multiple rows containing the exact same information), but as a rule each instance should possess some attribute in the base dataset—latent or otherwise—which makes it unique. If any attribute is missing from an instance, an invariant value (such as NULL) is assigned to connote the existence of missing data.

Oftentimes, tabular data will possess a relationship between one of the columns (the class variable) and the various other columns (the independent variables) which theoretically help explain change or variation in the dependent variable. Tasks which attempt to determine the quantitative nature of this relationship are characterized as supervised machine learning tasks. Though this relationship between the columns need not exist for tabular data, in this survey we will focus primarily on works in which the ultimate

task is one of supervised machine learning. Indeed, the problems caused by class imbalance which this paper will address will be those in which the dependent variable (alternatively, the "label") possesses a degree of class imbalance, and it is towards this class imbalance that the GAN will be applied.

With respect to the attributes, the various columns can display some degree of interdependence with one another, but the relationship between any two columns should not be deterministic or redundant. For example, a dataset related to cybersecurity traffic type should not contain attributes displaying both packet size in megabytes and packet size in kilobytes. Furthermore, attributes are characterized as possessing a certain data type, such as strings, integers, date values, binary values, etc. Generally, a given attribute will be invariant with respect to the data type it possesses; for example, an attribute meant to express integer values will not contain string values for a subset of records.

The use of deep learning methods such as GANs in tabular settings has been more muted than the implementation of these tools in computer vision tasks. Indeed, deep learning as a classification mechanism still takes a backseat in many tabular settings to non-deep learning techniques, such as tree ensemble methods like random forests and gradient boosting. In part, this is because deep learning has the most to offer in contexts where it can pattern intricate hierarchical representations of data, and where the local structure of data is conducive to convolutions and other such operations. Though image data and language data are strongly suited to the application of these mechanisms, tabular data generally does not exhibit such hierarchical or local structure.

Despite such relative disadvantages, research continues to find ways in which tabular datasets can benefit from deep learning frameworks. The benefits to using deep learning methods might be particularly noteworthy for large datasets, as is the case in the research of Haldar et al. [11], who outline how AirBnB uses non-deep learning methods for small problems and deep learning methods for large problems. Moreover, of particular importance for industry, deep learning methods allow for the training of prediction systems in an end-to-end fashion, such that non-tabular data can be integrated with tabular data with minimal alterations to the data pipeline. The development of Google's TabNet by Arik and Pfister [12] represents a significant step in the use of purely deep learning frameworks in tabular data, particularly with its "self-attention" mechanism which permits for a neural network to weight which elements of the input to focus on at any given time. Similarly, noting the prevalence of tree ensembles in tabular settings, the work of Popov et al. [13] adapts neural networks to this structure by implementing "soft" versions of decision tree splits.

Class imbalance

Researchers have long recognized the distorting effect class imbalance has had on machine learning predictions [1]. Class imbalance happens when the occurrence of the number of instances of one class or classes largely outnumbers the number of instances of other classes in a dataset. As most learners are designed with an assumption of relative parity of frequency amongst classes, the existence of class imbalance biases the learner towards classification of samples in the majority class. This bias presents a practical problem, as the minority class is often the class of interest with respect to classification. For example, a learner assigned the task of predicting whether patients have

benign (negative, majority class) or cancerous (positive, minority class) lesions will skew towards over-classifying the number of healthy patients (high false negative rate) rather than misdiagnosing healthy patients (high false positive rate). While the quantifiable definition of class imbalance varies with the research problem, He and Garcia [14] echo the viewpoint of many researchers that class imbalance spans from majority-to-minority class ratios of 100-to-1 or greater.

In general, academics have identified two approaches to correct class imbalance: algorithm-level methods and data-level methods. Broadly speaking, algorithm-level methods [15] alter the cost function within the learner—or otherwise, some wrapper generalization function—to prioritize the accurate classification of rare classes. Largely, we will ignore the algorithmic approach in this survey, as the generative capability of a GAN lends it towards data-level methods. These latter methods, as the name suggests, rely on resampling the dataset to minimize or eliminate the difference in the number of instances amongst classes.

Random Oversampling (ROS) and Random Undersampling (RUS) [16] are the two foundational yet effective methods of data sampling researchers have long used to treat class imbalance. ROS involves randomly replicating instances of the minority class, while RUS involves randomly removing instances of the majority class. RUS has the advantage of decreasing computational training time and avoids potentially harmful replication of data. However, undersampling methods possess the disadvantage of removing potentially valuable information. In contrast, oversampling preserves all instances in the original training set, at the expense of increased training time and possible overfitting [17]. At the extreme end of ROS implementation to correct for severe imbalance, the learner could simply generate a classification rule to encompass a single replicated instance.

Many variants of RUS and ROS exist that accentuate their stated advantages and minimize their stated caveats. A comprehensive survey of such methodologies can be found in the work of Van Hulse et al. [18], including a presentation of statistical significance with respect to the machine learning efficacy of each method. One of the most prominent modified oversampling frameworks introduced by Chawla et al. [19], Synthetic Minority Oversampling Technique (SMOTE) generates new minority instances via a nearest neighbor algorithm rather than randomly duplicating examples from the case library. A further modification to SMOTE yields Borderline-SMOTE [20], which prevents SMOTE from generating examples near class borders to maintain differentiable decision boundaries for classification. The system advocated by and Jo and Japkovicz [21] is another oversampling method which has garnered popularity. This approach, called Cluster-Based Oversampling (CBOS), views the true harm in class imbalance as its precipitation of small disjuncts, and addresses this degradation by identifying classes with k-means clustering and sampling these clusters iteratively. With respect to undersampling, many "intelligent" methods have been developed that address one-sided selection and minimize the removal of valuable information. Wilson's editing [22] is a nearest-neighbor based method that sharpens the decision boundary by removing possibly noisy labels. Alternatively, Kubat and Matwin [16] suggest methods to remove redundant samples with Tomek links [23].

In environments of deep learning, the question of class imbalance warrants special considerations. This is largely due to the way deep learning settings exacerbate

the problem posed by class imbalance, owing to the gradient method by which neural networks update their weights. As a consequence, training on an imbalanced dataset quickly reduces the error within the majority group in early iterations, while the misclassification of the minority group (or groups) increases and results in slow convergence [5]. Moreover, in scenarios with big data and severe data rarity with a very small number of samples of the minority class, the numerous parameters of a deep neural network could suffer from extreme overfitting by simply memorizing the minority instances.

In response to such idiosyncrasies, researchers have drawn inferences on the optimal treatment of class imbalance in these regimes. Given sufficient computational resources, oversampling has emerged as a preferable alternative to undersampling [6], particularly given the partitioning protocol of MapReduce processes [24]. Feature selection has proven an extremely valuable data preprocessing tool in boosting the efficacy of learning methods applied in imbalanced environments. There are feature selection methods which have been engineered with the existence of rarity in mind. Though an in-depth discussion of feature selection is beyond scope, Yin et al. [25] offer valuable insights into the confluence of class imbalance and feature selection. In cases of extremely severe rarity, where only a handful of instances of the minority class exist, reproducible learning patterns may be impossible to extract, regardless of sampling method [14]. A comprehensive summary of the treatment of class imbalance in deep learning settings has been published by Johnson and Khoshgoftaar [6].

The presence of class imbalance can have distorting effects on a machine learning model, biasing classifications in favor of the majority class. ROS and RUS have emerged as the two foundational data-level techniques to address class imbalance, and more advanced methods such as SMOTE, CBOS, and Wilson's Editing have likewise taken hold. Because of the unique challenges posed by classification in deep learning settings, researchers have adjusted the methodology by which they approach such prediction tasks. Namely, these adjustments include the prioritization of ROS over RUS wherever computationally feasible to avoid information loss, and pairing data sampling techniques with feature selection methods.

Generative adversarial networks

Since their introduction in [9], GANs have generated an intense amount of research interest. Observing industry and academic trends, LeCun has characterized GANs as "the most interesting idea in the last ten years in machine learning" [26]. As alluded to previously, GANs represent a competition between (generally) two neural networks, a Generator and a Discriminator. The Generator initiates the process by utilizing an input random noise vector to generate a data sample which looks as if it could have come from some base dataset. The Discriminator receives either this output from the Generator or an actual sample from the base dataset and must determine whether it is a forgery or a true data point. On the correctness (or lack thereof) of this characterization, the weights of the two neural networks are updated, and the process continues iteratively to the point where the zero-sum game reaches its Nash equilibrium: the Generator produces images with such fidelity that the Discriminator cannot tell the difference between a forgery and a true copy, tantamount to a guess.

The basic GAN structure explained above has been re-engineered to overcome many initial limitations, including vanishing gradients, mode collapse, and failure to converge. As noted by Wang et al. [27], these modifications tend to come in two forms: loss variants and architecture variants. With respect to the former, the original loss function in [9] often led to vanishing gradients. To combat this, many researchers have opted to replace the original loss function with more robust loss functions, such as Wasserstein loss. Architecture variants modify the neural networks constituent to the GAN (i.e., the Generator and the Discriminator). This includes refashioning the neural networks, generally the Generator, into a recurrent neural network, fully connected neural network, convolutional neural network, etc., or even adding other players (neural networks) into the min–max optimization game. The most leveraged alternative form of architecture-variant GAN is the Conditional GAN, or CGAN, discussed in further detail in the next section. CGANs condition the generation and discrimination of samples on the class labels, allowing for their appropriation into regimes of class imbalance.

Where traditional synthetic resampling methods (such as SMOTE) use local information to generate new instances, GANs use information from the overall class distribution. As instruments to correct class imbalance, GANs have generally been used in domains relevant to computer vision, particularly in medical imaging. As the research methodologies of pictorially oriented GAN applications are not the primary focus of this paper, a thorough review of GANs for imbalanced problems in computer vision can be found in the work of Sampath et al. [28]. In contrast, the implementation of GANs for data generation in tabular dataset problems has been more muted, even more so for such problems in regimes of class imbalance. To the extent that such work has been published, this research has generally dealt with anomaly detection in financial transactions or packet inspection in cybersecurity.

GANs are deep learning techniques that rely on the adversarial exchange between two neural network systems—Generators and Discriminators—to generate synthetic instances of a dataset. Since the introduction of the basic GAN framework, the development of the CGAN methodology and other variant architectures have accommodated GANs for their incorporation into machine learning tasks that suffer from class imbalance. Though the application of GANs toward this end has primarily been reserved for computer vision problems, GANs are increasingly being implemented in tabular domains, specifically those related to cybersecurity and financial transactions.

Basic GAN frameworks

GANs in their original form possessed no inherent mechanism to specify the desired class to generate for an instance, and thus their implementation as tools to correct class imbalance was limited. As a result, researchers have focused on modifying the GAN architecture to allow for class conditioning to help generate specified minority classes, most notably in the form of CGAN and BAGAN. These two imbalance-sensitive variants have themselves served as the basic frameworks for subsequent research efforts, which include the incorporation of encoder-decoder modules, the addition of other components to the Generator-Discriminator interplay, and the modification of the loss function in the Generator and Discriminator.

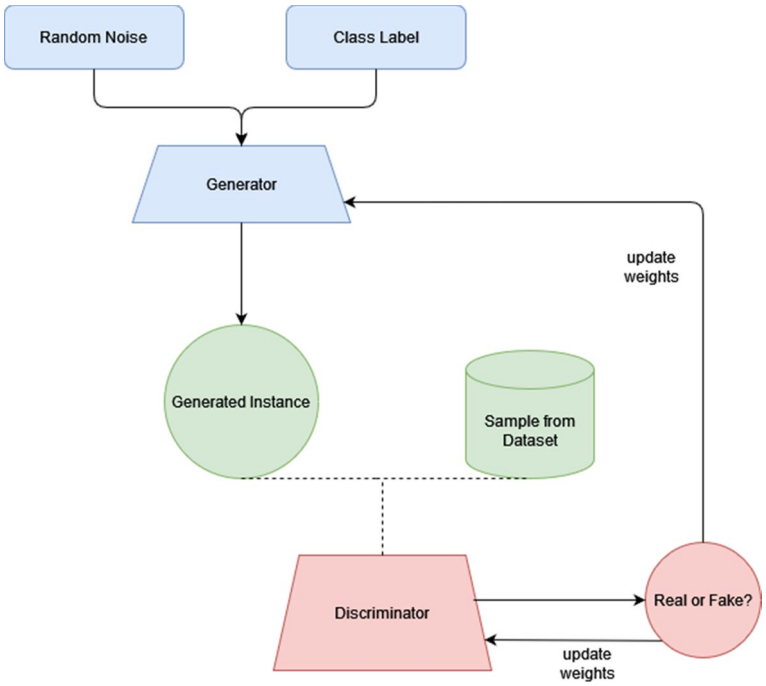

Fig. 1 CGAN architecture

Osindero and Mezri [29] laid the groundwork for future class conditioned GANs in their development of CGAN. Though it only represents a slight modification of the basic GAN architecture, the intuition to condition the input space of the Generator and Discriminator with class labels from the dataset marked the first allowance for disproportionate generation of minority classes. Douzas and Bacao [30] applied CGAN on 12 different imbalanced datasets from the UCI Machine Learning Repository, as well as 10 other datasets from the Python library Scikit-learn [31], and baselined it against Random Oversampling, SMOTE [19], Borderline SMOTE [20], ADASYN [32], and Cluster SMOTE using a number of common learners. CGAN possessed the highest mean composite ranking with respect to AUC, geometric mean, and F1 score, with statistically significant results relative to other methods at the 95% confidence level. A diagram of the CGAN architecture is given in Fig. 1.

Another seminal approach can be found in the research of Odena et al. [29] and their auxiliary classifier GAN (AC-GAN). This framework supplements the Discriminator with a decoder network to output class labels for data, rather than binary assignment of real or fake. Antoniou et al. [30] spell out an early and well-cited approach to data augmentation via GANs, though their concern is not directly related to class imbalance. Because of its independence of class, their DAGAN model can be applied in extremely low data regimes, even to unseen classes.

The work of Mariani et al. [31] is perhaps the first to explicitly define the scope of GANs as reformers of class imbalance. The insight offered by the authors is that in a game theoretical zero-sum competition between the Generator and the Discriminator where a minority class exists, and where—for the sake of argument—the Generator has learned to generate realistic minority class images, the Discriminator will encounter a

rare but real sample of a minority class and classify it as fake. In keeping with its optimization function, the Discriminator will classify thusly for all minority instances. The Generator, in turn, will be rewarded for "fooling" the Discriminator, and subsequently begin to output images that are not representative of the minority class. Consequentially, the researchers developed the Balancing GAN (BAGAN) framework to prevent this potentially destructive, uninformative interplay. The BAGAN methodology, since enhanced or baselined against with frequency, proposed only rewarding the Generator for generating instances which both fool the Discriminator and can be correctly assigned to the desired class label. Thus, the architecture calls for initializing the Generator with an autoencoder to help avoid mode collapse, and a decoder to translate latent features into a probability of whether the image in question is fake, or a probability the image corresponds to a class. Applied to four datasets, BAGAN generally outperforms a vanilla GAN and AC-GAN with respect to accuracy. Moreover, BAGAN outperforms these baseline approaches with respect to variability in image quality, as measured by structural image similarity (SSIM).

One of first the end-to-end GAN frameworks to address class imbalance can be found in Generative Adversarial Minority Oversampling (GAMO), developed by Mullick et al. [32]. Here, the researchers suggested a three-player adversarial game inspired by Mariani et al. [31], whereby the Generator simultaneously tries to fool the Discriminator with real or fake instances and a classifier with convex combinations of minority instances. These instances are drawn from and generated at near the boundary lines of respective classes, rather than from the center of class distributions as with previous approaches [33]. Moreover, the Generator is constrained to synthesize instances that do not fall outside of the observed distribution of respective minority classes. The researchers test their GAMO infrastructure on seven datasets, with class distribution created where necessary by randomly removing majority samples. GAMO generally outperformed predecessor methods with respect to Average Class Specific Accuracy (ACSA) and Geometric Mean.

Cybersecurity

Cybersecurity is a realm where the opportunities for rich, synthetic data are rife, and one which demonstrates the appropriation of GANs as class balancers need not necessarily be solely for imaging tasks. Though there are many machine learning tasks that fall under the broad umbrella of cybersecurity, in general most problems in the domain relate to network traffic classification at the packet level. Practitioners are interested in determining whether a classification system can detect malicious packet streams and, by extension, network attacks. However, since instances of malicious packet streams are comparatively rare events, these datasets tend to suffer from class imbalance, which can make the design of effective intrusion detection systems difficult [34]. Khoshgoftaar and Leevy [35] provide a survey of such approaches, many of which are in environments of deep learning, though the use of generative methods at the time of publication had been too infrequent to include. In general, most researchers still rely on traditional, non-generative methods of data sampling to correct for class imbalance, though publications investigating the feasibility of GANs have garnered increased attention in the past two years. The papers reviewed in this section use generative methods to create instances of rare, malicious packet traffic,

and baseline the final classification results against datasets generated by other oversampling techniques.

Vu et al. [36] are the first to explore the potential for GANs to alleviate class imbalance in this domain, seeking to procure classifications in a binary setting as SSH traffic (majority class) and non-SSH traffic (minority class). Using the AC-GAN methodology and the Network Information Management and Security (NIMS) dataset [37], this work baselines its generative method against a SMOTE-augmented dataset and a BalanceCascade-augmented dataset [38]. Though the AC-GAN took, on average, 4 to 5 times longer to train, it yielded a higher accuracy, AUC, and F1 score than all other baseline methods.

Lee and Park [39] also leveraged GAN-enhanced datasets for anomaly detection tasks. Here, the team was particularly interested in the performance of certain traffic types. The researchers used the CICIDS 2017 dataset [40] to construct a basic GAN and augment the various classes in proportion to their rarity in the base dataset. This is in contrast to many prior and subsequent approaches, which use CGAN or some such variant to "fill in the gaps" with respect to class imbalance, rather than selectively resampling in a piecemeal class-wise manner; in total, ten thousand additional rare instances were generated. Using Random Forest as a classifier, the GAN-based approach outperformed no treatment with respect to accuracy (99.83% vs. 99.19%), precision (98.63% vs. 98.2%), recall (92.76% vs. 83.79%), and F1 score (95.04% vs. 87.79%). The performance gains were particularly noteworthy for the three classes with the greatest rarity: Bot, Infiltration, and Heartbleed.

Wang et al. [41] also sought to apply GANs to the problem of encrypted traffic classification, particularly with an eye towards generating classes which have very sparse "natural" occurrences in packet traffic and are therefore harder to classify. The team adopted the architecture of AC-GAN using the ISCX2012 corpus [42]. After data generation, the research team randomly reduced the size of the majority class and adjoined the minority class to bring the training data to parity. With no treatment and Random Oversampling as baseline methods, the various datasets were trained using a generic five-layer Multilayer Perceptron. The researchers found that FlowGAN generated data which trained classifiers with accuracy (0.991), precision (0.9911), recall (0.991), and F1 score higher than any of its counterparts. Moreover, it was noted that in contrast to no treatment and Random Oversampling, FlowGAN did a commendable job in discerning known similar traffic types.

The work in [41] is reprised and enhanced by Wang et al. [43], featuring much of the same research team and once more utilizing GANs to rebalance network classification traffic datasets. The authors used a repurposed form of conditional GANs dubbed PacketGAN. In addition to initialized random input noise, the team also fed the label of traffic types with one-hot encoding into the GAN at multiplied vector lengths of 1480 to generate class-balanced synthetic samples. PacketGAN was compared against baselines of Random Oversampling, SMOTE, and vanilla GAN data synthesis methods, as well as an untreated imbalanced dataset. Generated on the ISCX2012 [42] and USTC-TFC2016 [44] datasets, and classified using MLPs, Sparse Autoencoders (SAE), and CNNs, PacketGAN yielded classifications with accuracy (0.995), precision (0.994), recall (0.996), and F1 score (0.995) higher than any of its counterparts.

In [45], Yilmaz et al. used GANs to rebalance the UGR'16 dataset [46] to create an intrusion detection system, identifying attack classes such as 'anomaly-sshscan', 'nerishbotnet', 'blacklist', 'anomaly-spam', 'anomaly-sshscan', 'dos', 'scan 11', and 'scan 44'. To prepare the data for consumption by a GAN, the research team mapped source and destination IP addresses to unique numeric values, and normalized all other attributes, resulting in a dataset at a weekly time interval with 12 features and over one billion instances. The authors optimized a Generator and Discriminator with five hidden layers each and a ReLU activation function with learning rates of 0.02 and 0.0025, respectively. Utilizing a sixty percent-forty percent train-test split and baselined against an untreated dataset, the team determined the data augmentation abilities provided by the GAN allowed for significant classificatory performance improvements for the minority classes. After rebalancing, the recall, precision, and F1 score for all classes were above 99%, whereas prior to rebalancing the same metrics for all minority classes were below 67%. The greatest gains came for the 'dos', 'scan 44', and 'blacklist' classes. The authors noted the hardware and software limitations which fragmented their data pipeline, as well as the data preprocessing methods necessitated to prepare the data for ingestion by a neural network.

Belenko et al. [47] orient the use of GANs for intrusion detection in machine-to-machine (m2m) communication networks, increasingly common with the emergence of the Internet of Things (IoT). However, rather than evaluating the generated data against baselines by machine learning efficacy, the researchers offer sequential narration of the generation of a viable synthetic dataset. The dataset in question contains eight ratio numerical attributes, and the GAN architecture to which the authors devoted the greatest amount of focus was the CGAN. After approximately 6000 training iterations, the researchers determined the Generator and Discriminator curves converged with respect to the loss function.

Table 1 provides a summary of the average results of the experiments discussed in this section. The results are broken down by research paper, whether the method in question was experimental (novel) or merely baseline, and the specific sampling method used. For each paper and performance metric, top performers are bolded and italicized. By exposition of this aggregation, an experimental method outperformed baseline methods in all but two instances, both occurring in [41]. Here the use of no treatment results in an optimal precision, and the use of oversampling results in an optimal recall. Across all other methods, the many devised experimental methods outperform their baseline counterparts in AUC, balanced accuracy, F1 score, precision, or recall.

Financial transactions

The unprecedented computerization of the global financial system has yielded vast troves of data with which researchers and practitioners can accomplish innumerable analytical tasks. One such task is fraud detection, whereby a model tries to determine which transactions among a set of transactions are fraudulent. Another example is credit loan application, whereby an auditor uses transaction payment history and other information about the applicant to make a decision on creditworthiness. However, both of these assignments are beset by the problems of class imbalance; the majority of transactions are perfectly legitimate [48], and—in the retail contexts in

Table 1 Average results of experiments related to cybersecurity, aggregated by paper, novel vs. baseline, and sampling method

Paper - Novel vs Baseline - Sampling Method	AUC	Balanced Accuracy	F1 Score	Precision	Recall
A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification					
Novel					
ACGAN	88.42%	99.80%	86.67%		
BalanceCascade	***95.10%***	***99.90%***	***95.10%***		
Baseline					
no treatment		99.49%	82.92%		
SMOTE	94.12%	99.20%	94.94%		
FLOWGAN:Unbalanced network encrypted traffic identification method based on GAN					
Novel					
ACGAN		***99.10%***	***99.10%***	97.99%	89.95%
Baseline					
no treatment		89.95%	89.68%	***99.11%***	97.94%
Oversampling		97.94%	97.96%	90.00%	***99.10%***
GAN-based imbalanced data intrusion detection system					
Novel					
GAN RF		***99.83%***	***95.04%***	***98.68%***	***92.76%***
Baseline					
RF		99.19%	87.79%	98.20%	83.79%
SMOTE		99.51%	88.16%	88.97%	87.51%
PacketCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN					
Novel					
ACGAN		***99.51%***	***99.47%***	***99.36%***	***99.58%***
Baseline					
GAN		97.66%	97.66%	97.66%	97.67%
No treatment		97.97%	97.66%	97.59%	97.75%
Oversampling		98.89%	98.91%	98.92%	98.89%
SMOTE		97.69%	97.10%	97.51%	97.89%

Average results are displayed by evaluation metric, and top performers are bolded and italicized

which they are studied—most loan applicants end up repaying their loans. As a result, many researchers have turned to data sampling techniques to help remedy class imbalance. The use of GAN-based methods remains limited in its scope of spread, but the papers discussed in this section have utilized GANs to generate instances of the rare, positive class and have baselined these instances against traditional data sampling techniques, with very favorable machine learning efficacy results.

Fiore et al. [49] impose GAN generation on data from Dal Pozzolo et al. [50] in a regime of extreme imbalance to train a classifier to identify instances of fraudulent credit card transaction (0.172% occurrence of fraudulent transactions). Where most methods discussed have used a type of conditional GAN for the generation of minority class instances, the researchers in this work instead train the GAN only on the

subset of positive cases in the dataset (315 records), later merging the artificial fully-minority dataset back into the original training corpus. Baselined against SMOTE and evaluated by precision, F1 score, and accuracy, [49] find that effectively doubling the number of fraudulent cases via synthetic injection results in the optimal GAN. Though the GAN methodology outperformed the SMOTE methodology, the margin was not statistically significant.

In contrast to prevailing methods, Lei et al. [51] bypass the clear separation of the Generator and the learner and combine the two in a binary classification setting of credit scoring into a framework they term IGAFN. Rejecting the tradition of what they call "GAN-separated methods", the researchers instead implemented an architecture they term a "fusion module" (the primary contribution of the paper); that is, the harmonization of a customer's attribute level data (age, income, etc.) and time series data (history of credit payments) outputted a sample to the Generator. In the vein of SGAN [52], the Generator must subsequently determine whether the sample is fake, real positive, or real negative. The generated data is then used to augment the original dataset, and the Generator itself is used in a transfer learning manner for subsequent classification. Testing this method against a number of other baseline methods (including SVM with SMOTE augmentation and a traditional separated GAN), [51] found that IGAFN outperformed all other methods with respect to accuracy (85.65%), AUC (73.57%), and F1 score (0.6101) using the Credit Card Clients Dataset [53] as the base corpus. Likewise, using the Nigeria Credit Risk Prediction Data[1], the research team again attained the best results with the IGAFN methodology (83.99%, 71.12%, and 0.5851). Subsequently, the authors varied the proportion of generated minority samples they actually injected into the database and found a proportion of 1 generated positive sample for every 2 real positive samples yielded learners with the highest AUC.

Recently, Engelmann and Lessmann [54] also examined the ability of GANs to generate data in a structured (tabular) rather than unstructured (image) context, specifically in the field of credit scoring. Like Quintana and Miller [55], these researchers sought to generate and use both continuous and categorical explanatory variables. The authors opted for a Wasserstein GAN [56] architecture, with adjustments such as using the Gumbel-softmax activation function [57] in combination with embedding layers [58] to model discrete numerical variables, and min–max scaling paired with the addition of Gaussian noise data to avoid Discriminator detection of a trivial pattern ("number of loyalty points", for example, which in the real dataset only appears in increments of ten). The treatments applied herein generate examples whose individual variable distributions are believably close to the variable distributions of their real counterparts.

The use of a Wasserstein GAN represents an important shift in the optimization of the loss function and the role of Discriminator. Now the Discriminator acts like a"critic", motivated by the intuition that the Generator should minimize the distance between the distribution in generated samples and the distribution of actual data in the training dataset [56]. This circumvents the limitations of a traditional Discriminator, wherein a fully-trained network may cease to supply useful gradient information in iterative calibration

[1] https://www.kaggle.com/c/data-science-nigeria-credit-risk-prediction.

of the Generator, and seek to only fool the Discriminator rather than generate data samples faithful to the original distribution. Using Wasserstein loss can result in a stabler training process with more realistic synthetic instances.

The use of the Wasserstein loss function, combined with the above data treatments, yields a resulting framework called cWGAN, and is implemented on seven different datasets and benchmarked against no treatment, Random Oversampling, SMOTE, SMOTE-Nominal Continuous, ADASYN, and Borderline SMOTE using five different learners. Evaluated on AUC, AUC-PRC, and the Brier Score, cWGAN fared favorably, outperforming SMOTE variants on five of the seven datasets. However, [54] note some areas of suboptimal performance, particularly on datasets with linearly separable classes. Because the harmful effect of class imbalance may be less severe on such datasets [59], more advanced methods such as cWGAN may not be necessitated. However, on the two "complex", strongly non-linear datasets, cWGAN was the clear winner. Ablation studies further conclude that the general outperformance of cWGAN is due to the specific GAN architecture the authors engineered.

Table 2 offers a synopsis of the average results of the statistical experiments discussed above. As in the previous section, the values are grouped by research paper, the nature of the sampling methodology (novel/experimental or baseline), and the sampling method used, if any, for the correction of class imbalance. Top performers are bolded and italicized for each paper and performance metric. Surprisingly, the absence of data sampling results in optimal machine learning efficacy in [54] (that is, the ordinally lowest rank). Moreover, in [49], the use of SMOTE outperforms a GAN-based approach with respect to specificity (100% vs. 99.99%), though the difference by some may be considered to be negligible. In [51], the use of the fusion-module equipped IGAFN attains a higher average AUC, balanced accuracy, and F1 score than any of the baseline methods.

Other disciplines

The abilities of GANs as treatments to class imbalance have been utilized outside of the domains of cybersecurity and financial transactions. Some of these research endeavors likewise target specific industry applications such as human comfort, while others are domain-agnostic and aim for interdisciplinary implementation on a wide variety of datasets. A few papers examine the impact of modifying the Discriminator—Generator interplay by adding other components, the impact of modifying the loss functions, or the impact of adopting a hybrid method which emphasizes sample weighting rather than data generation. Still others attempt to answer other questions relevant to proper implementation of GANs in tabular settings, without a specific aim of data generation or machine learning efficacy.

The work of Wang et al. [60] represents an early exploration into the alleviation of class imbalance via GANs, as well as the use of GANs in fine-grained classifications. In this research, the authors leverage the GAN methodology to generate data of rare plankton species for the purposes of classification. Designating their model as CGAN-Plankton, the researchers modified the CGAN architecture to generate rare instances of plankton from the WHOI-Plankton dataset [61], and compared the classifications of this generated dataset against various CNN baselines which received no treatments for class imbalance. The research team found that CGAN-Plankton yielded the highest accuracy

Table 2 Average results of experiments related to financial transactions, aggregated by paper, novel vs. baseline, and sampling method

Paper - Novel vs Baseline - Sampling Method	AUC	Balanced Accuracy	F1 Score	Ranking	Recall	Specificity
Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning						
Novel						
CWGAN				3.15		
Baseline						
ADAYSN				5.37		
B-SMOTE				4.03		
None				***2.32***		
Random				3.48		
SMOTE				3.90		
SMOTE-ENC				5.73		
SMOTENC				5.17		
Generative adversarial fusion network for class imbalance credit scoring						
Novel						
IGAFN	***71.08%***	***83.60%***	***57.98%***			
Baseline						
CFN	65.64%	76.05%	47.94%			
GAN	70.07%	79.39%	54.43%			
No treatment	63.55%	79.35%	42.12%			
SMOTE	65.38%	70.98%	43.27%			
Using generative adversarial networks for improving classification effectiveness in credit card fraud detection						
Novel						
GAN					***71.94%***	99.99%
Baseline						
SMOTE					70.60%	***100.00%***

Average results are displayed by evaluation metric, and top performers are bolded and italicized

and F1 scores, with a marked advantage in those metrics localized for rare classes. The authors likewise noted the fidelity of the generated instances to actual instances. The modifications to GAN architecture are fairly innovative given incorporation of a classification adversarial network in the interplay between the Generator and the Discriminator, though most future research in multiple domains would diverge from their decision to generate synthetic data based off only small class instances.

Another mostly unexplored application of GANs in imbalanced data settings in the realm of tabular datasets comes relating to human sentiment. Quintana and Miller [55] attempt to remedy class imbalance in a human comfort dataset [62], a dataset inquiring of participants the satisfaction with their living environments which contains a sizeable majority of "0" (neutral) labels. The research duo examines the performance of the Tabular-GAN framework developed by Xu and Veeramachaneni [63], as well as no treatment, GANCorr, and a basic GAN as baselines, all with a 70–30

train-test split and with KNN, Naïve Bayesian, and SVM learners. For numerical features, Tabular-GAN applies mode-specific normalization using a Gaussian Mixture Model (GMM), thereby addressing the problems posed by multimodal distributions. To prepare categorical variables for ingestion into a neural network, the model calls for implementation of the Gumbel softmax function, which one-hot encodes a discrete variable, adds noise, and then normalizes the attribute. The results of Tabular-GAN with respect to F1 score were mixed, especially when the number of generated samples used in training surpassed the number of real samples used in training. However, [55] note the maintenance of baseline performance indicates that Tabular-GAN was able to adequately capture the relationships between features and suggest that more nuanced deployments of Tabular-GAN in scenarios with more available data may yield superior results.

To that end, the same core of researchers in [55] expanded the scope of their study in [64], including more training datasets and hyper-parametrizing a bespoke model, called ComfortGAN. ComfortGAN relies on the CGAN architecture with Wasserstein loss and a gradient penalty. For both the labels and the categorical fields, the researchers converted them via one-hot encoding, added uniform noise, and renormalized them, like the process followed by Xu and Veeramachaneni [63]. Continuous features were scaled between -1 and 1. The Generator consisted of five fully connected layers with a ReLU activation function, while the Discriminator possessed a Leaky ReLu activation function. Experiments were evaluated by a Euclidean distance measurement for variability of samples and diversity of generated samples with respect to the training set, as well as F1 score for machine learning efficacy. Baselined against no treatment, SMOTE, ADASYN, TableGAN, and CTGAN, ComfortGAN outperformed all of its competitors across the board. Subsequently, the group reprised its experiments, now adjusting the datasets to consolidate the number of predictable classes down to 3. Once again, ComfortGAN retained its statistical advantage, but the margin was less pronounced. This led [64] to speculate that for the research problem in question, the complexity and computational load of a GAN-based approach may not be justified, particularly when, as in the case with human comfort survey data, classes can be combined, and greater emphasis can be placed on fine-tuning the machine learning models themselves rather than data augmentation.

The work of Dos Santos Tanakha et al. [65] represents another example of the application of GANs on tabular datasets. The authors here opted for an exploratory and general approach to evaluating the rarity rectifying properties of GANs, with an emphasis on exploring GANs with different architectures and layers of complexity. The researchers deployed six different GANs as base Generators, ranging from a GAN with one hidden layer with 128 nodes, to a GAN with three hidden layers and as many as 1,024 different nodes in a given layer. The group also used a variety of well-drawn corpora, including the Pima Indians Diabetes Database [66], the Breast Cancer Wisconsin Dataset [67], and the Credit Card Fraud Detection Dataset [68] on all of these architectures, all of which suffer from varying degrees of class imbalance. Though decision tree learners trained on purely synthetic data had higher accuracy than learners trained on the original datasets, the classificatory results were less conclusive with respect to using GANs to rebalance these datasets. Compared against

baseline methods of SMOTE and ADASYN, the GAN-based methods yielded classifiers which were no more performant on rebalanced data with respect to accuracy, recall, and precision.

Deepshikha and Naman [69] enhance the traditional GAN architecture by adding another adversarial network to compete with the Generator and the Discriminator, called the Classifier. In addition to the standard competition with the Discriminator, the Generator also competes in a min–max game the with Classifier to fool the Classifier with respect to class assignment (rather than real versus fake). The motivation for this maneuver is born of the desire to generate samples in the "convex hull" [70] of the training classes. The resulting architecture (called Polarity GAN, or PGAN) ensures the Generator creates data points close to the decision boundaries between classes, thereby generating more challenging examples. To ensure robustness, [69] evaluated PGAN and the classifications resulting from its dataset against six different baseline methods (including AC-GAN, BAGAN, and WGAN-GP) on nine different corpora. PGAN outperformed its competitors on eight of nine datasets with respect to both F1 score and average accuracy, and consistently performed optimally in alleviating the uncertainty surrounding the determinations of the Discriminator (broadly defined as uncertainty sampling, as measured by Least Confidence, Margin of Confidence, Ratio of Confidence, and Entropy).

To address bimodal distributions, Xu et al. [71] utilize a variational Gaussian mixture model (or VGM, as introduced by Bishop [72]) to fit a mixture of Gaussian curves. Tabular data suffers from the potential drawback of multimodal and non-Gaussian distributions in continuous numerical columns, leading to the problem of vanishing gradients. This means traditional methods such as max–min normalization cannot be applied to such attributes. To overcome this and make tabular datasets copasetic to a GAN, the authors devised the concept of mode-specific normalization, whereby the various modes in a multimodal distribution are identified, sampled with a random probability, and then normalized. Tabular data can also display imbalance in discrete and categorical attributes, a problem relatively unexplored in the context of generative methods and deep learning. The researchers addressed this by generating synthetic rows conditioned on a given discrete column, whose value is sampled by the log-frequency of occurrence so as to minimize the impact of rarity. Of equal interest, the authors adapted the same preprocessing methods to a variational autoencoder with a lower-bound loss function [73]. The resulting architectures—dubbed CTGAN for the GAN and TVAE for the autoencoder—are trained and validated against seven simulated datasets, generated from Grid and Ring Gaussian mixture oracles [74] with random offset, and six popular datasets commonly extracted from the UCI machine learning repository, including the credit risk dataset from Kaggle and a binarized form of the MNIST dataset. Evaluated on likelihood fitness metric for simulated datasets and machine learning efficacy on real datasets, and benchmarked against a number of comparable Bayesian network and GAN architectures, [71] found the TVAE outperformed all other methods, including CTGAN, on five of six dataset-evaluation combination aggregations. The lone instance of optimal performance for CTGAN came on Gaussian mixture simulated data using the likelihood fitness metric; however, this does not establish a dominance of VAEs over GANs, as the authors

point out the latter can achieve differential privacy [75] with greater facility than the former.

Ren et al. [76] propose a domain-agnostic approach to improving the machine learning efficacy of GANs. Specifically, the team suggests infusing the WGAN architecture with a class label vector weighted by entropy to help characterize the data imbalance among the classes. This entropy-weighted vector is then concatenated to the original feature vector and fed into the Generator. The resulting chassis, dubbed Entropy-based Wasserstein Generative Adversarial Network (EWGAN), improves performance in regimes of high data imbalance. As proof, the researchers deployed EWGAN on the Vowel0 and Page-blocks0 datasets (two relatively small benchmark corpora with 2.4% and 5.6% class imbalance ratios, respectively) against baselines of no treatment, K-means SMOTE, WGAN, and CGAN with a support vector machine as universal classifier. On both datasets, EWGAN outperformed all other data augmentation methods with respect to accuracy.

Montahaei et al. [77] did not seek to generate synthetic examples, but rather utilized the adversarial process to assign weights to the negative (majority) class via the Generator. Higher weights are assigned to negative instances that lie near the decision boundary, thereby emphasizing the importance of informative negative samples. These re-weighted samples were then fed to the Discriminator, which doubled as a classifier and outputted a positive class probability. In this sense, the research method resembles a hybrid between a data-level approach and algorithm-level approach. Evaluated on five different datasets with a logistic regression learner, [77]'s method (notated as Adversarially Re-weighting for Imbalanced Classification, or ARIC) outperformed other data augmentation methods—including SMOTE and ADASYN—on three of the five datasets with respect to accuracy, AUC, precision, and F1 score. A limitation of the work thus far, however, is its restricted applicability to binary classification settings. Likewise, [78] did not concern itself with data generation per se but apply GANs towards healthcare in a regime of sizeable class imbalance.

Where most research has focused on modifying GAN architecture to achieve optimal results in class imbalanced settings, Mizra et al. [79] posed a distinct yet equally important question: given optimized performance on a desired evaluation metric, what data augmentation method and proportion of synthetic sample injection should be used? The resulting framework, termed Model-Metric Mapper methodology, or MMM, can conversely offer a procedural and hierarchical approach to guide the practitioner toward proper model selection based on desired evaluation metric. The authors identified three broad classes of sampling techniques: reductives, synthetics, and generatives, the last of which encompasses GAN methodologies, variational autoencoders (VAE), and restricted Boltzmann machines (RBM). Training models on datasets from multiple domains, the authors propose GANs are best employed to resample the data distribution to nearly 50% class imbalance (i.e., resampled to near parity), with AUC, geometric mean, or balanced accuracy as the evaluation metric.

Taken together, these various research works demonstrate how GANs can be improved to treat class imbalance. The adjustments discussed herein include the use of Wasserstein loss in training the GAN components, the inclusion of an additional component (the Classifier) to generate more challenging examples, the use of autoencoders on the

Table 3 Count of frequency of metadata elements of methodologies in papers studied in this survey. Includes classifier, evaluation metric, dataset, year, and baseline method

Element	Value	Count	Element	Value	Count
Classifier	CNN	4	Dataset	MNIST	4
	DT	3		CIFAR-10	3
	KNN	3		GM Sim	2
	SVM	3		ISCX	2
	LR	2		Blog Catalog	2
	MLP	2		CelebA	2
	GBC	2		CICIDS 2017	2
	RF	2		SVHN	2
	SAE	1		UCI Kaggle	2
	NB	1		Wikipedia	2
Evaluation Metric	Balanced Accuracy	13	Year	2015	1
	F1 Score	11		2017	5
	Precision	7		2018	3
	Recall	7		2019	8
	AUC	5		2020	2
	Geometric Mean	3	Baseline Method	SMOTE	10
	Ranking	2		no treatment	8
	Specificity	2		Oversampling	4
	SSIM	2		VAE	4
	Likelihood fitness metric	1		ADASYN	3
	FID	1		Kmeans-SMOTE	2
	Euclidean Distance	1		RBM	2

latent space, and a broad consideration of the regimes, evaluation metrics, and learners in which the use of GANs is appropriate. Moreover, in sequential research instances GANs were also applied to correct class imbalance on human comfort data. However, the results here and in other research works suggest that the computational complexity of GANs is often not justified by the positive yet marginal lift they provide to machine learning classifications, especially when the base dataset in question can be simplified to make it compatible with more traditional data sampling and machine learning methods.

Meta-analysis of methodologies

In this section, we undertake a statistical analysis of the configurations associated with the various experiments mentioned above. The data in the analysis below was compiled by manually extracting metadata related to the results of the experiments of the research papers thus far recapitulated. This included extracting information produced in charts, tables, and figures of the respective papers, such as the evaluation metric, the value of the metric in question, the learner used in classification, the data sampling method, etc. This also includes manual parsing of the methodology section for each paper, making note of information such as the GAN architecture, the dataset used, the number of classes, the imbalance ratio, and—of paramount importance for the successive analysis—whether the method in question is to be considered "novel" (i.e., experimental, or relevant to the GAN innovation the author is introducing) or primarily baseline.

Table 3 provides an overview of some of the elements of the methodologies deployed by the papers studied herein. Each grouped count represents a unique implementation of that instance within a paper. Four of the papers use convolutional neural networks as classifiers, and three papers used decision trees, k-nearest neighbors, and support vector machines.[2] The most popular evaluation metrics for the various research works were balanced accuracy and F1 score, appearing in 13 and 11 different papers, respectively. All datasets with representation in two or more papers are shown in Table 3; the MNIST and CIFAR-10 datasets are used with great frequency by the "foundational" GAN papers explored in this survey, while the ISCX and CICIDS 2017—with two appearances apiece—are popular corpora for intrusion detection tasks. As noted, all research explored in this synopsis has been conducted within the past six years, with over half of the papers of interest published in the previous two years. With ten paper-instances of use, SMOTE registers as the most common baseline method used in sampling, followed by oversampling with four; the use of baseline methods untampered by sampling procedures occurs in eight separate papers.

In general, research which has focused on the use of GANs in tabular data has adopted one of three broader GAN structures: AnoGAN, CGAN, or CGAN with added components. The AnoGAN method is based on the work of Schlengl et al. [78], and applies the generative power of GANs only to the positive classes; hence the base dataset is "pre-filtered" for instances of the minority class. Using CGAN, the entire dataset is retained, but the class label is fed into the Generator to condition the class of the generated instance. From CGAN, there are a number of architectural modifications which have introduced other players into the adversarial contest between the Generator and Discriminator, most notably the Classifier to assign predicted classes to instances during the GAN process, or the inclusion of an autoencoder to translate the latent space. These we may consider as "CGAN with other modules"; BAGAN is one prominent example of this lineage. This delineation is presented in Fig. 2.

Table 4 provides this categorization for the thirteen papers related to tabular data which explicitly stated the implemented architecture. As evident, CGAN was the primary base architecture for nine of the thirteen papers, used in cybersecurity, financial transaction, and miscellaneous areas of research. The method has proven particularly popular in the realm of cybersecurity, as all of the examined research which stated its architecture in this realm deployed some modified form of CGAN. AnoGAN and CGAN with other modules each had two use cases, with both appearing in the literature pertinent to financial transactions.

The use of AnoGAN in the domain of financial transactions but not cybersecurity may speak to the specific implementation of AnoGAN in a domain of severe rarity. As suggested by Johnson and Khoshgoftaar [6], the ability of a deep learning scheme to recognize patterns may not be solely affected by the ratio of class imbalance, but also by the raw total of minority samples. Even with severe class imbalance, if the quantity of minority cases is sufficient, a pattern may be detected; however, in analogous case

[2] Note that the figures from this breakdown do not include categorizations for which the respective authors have not explicitly defined. For example, [58] use Random Forest, logistic regression, gradient boosting, KNN, and decision tree classifier, but do not break out the reporting of their results at the classifier level, and therefore are not represented granularly in this tabular analysis.

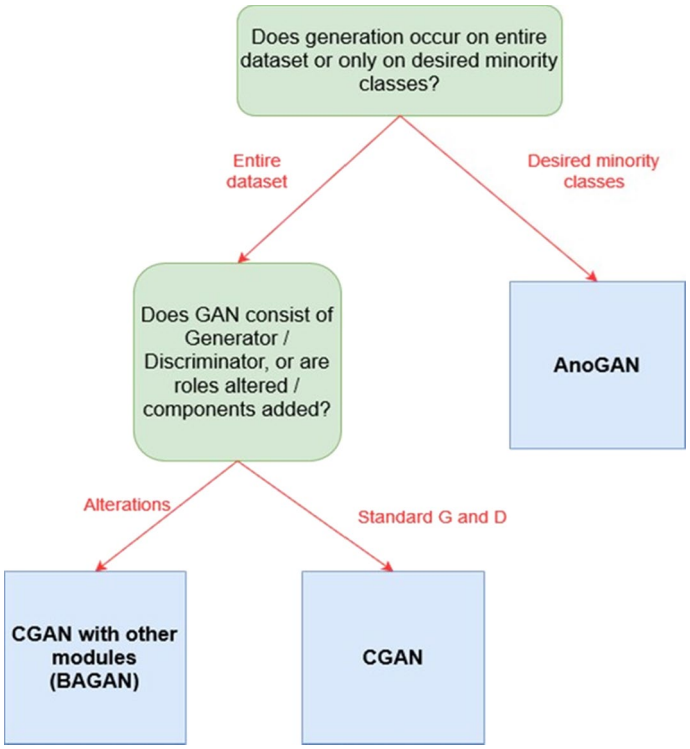

Fig. 2 Delineation of common GAN methodologies for tabular data

Table 4 GAN architectures of research works related to tabular data where explicitly stated, broken down by research area

Architecture	Research area	Research works
AnoGAN	Financial Transactions	Fiore et al. [49]
	Other Disciplines	Wang et al. [60]
CGAN	Cybersecurity	Belenko et al. [47]
		Lee and Park [39]
		Wang et al. [44]
		Wang et al. [43]
		Vu et al. [36]
	Financial Transactions	Engelmann and Lessmann [54]
	Other Disciplines	Quintana and Miller [55]
		Quintana and Miller [64]
		Dos Santos Tanakha et al. [65]
CGAN with other modules	Financial Transactions	Lei et al. [51]
	Other Disciplines	Deepshikha and Naman [69]

with less severe imbalance, if there are nevertheless significantly fewer instances of the minority class, machine learning algorithms may fail to detect a pattern. Likewise, in situations with severe class imbalance, the architecture provided by AnoGAN (which effectively filters and feeds the GAN networks with only minority examples) may be preferrable to the architecture of CGAN. This was precisely the situation in the work of Fiore et al. [49], where there were only 315 available positive instances.

Table 5 Novel vs. baseline results, aggregated at the level of different evaluation metrics for the various papers

Paper	Metric	Novel	Baseline	Baseline or Novel?
A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffi	AUC	0.8926	0.9412	baseline
A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffi	Balanced Accuracy	0.9985	0.9945	novel
A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffi	F1 Score	0.8836	0.8592	novel
Adversarial Classifier for Imbalanced Problems	AUC	0.9674	0.8809	novel
Adversarial Classifier for Imbalanced Problems	Balanced Accuracy	0.9619	0.9150	novel
Adversarial Classifier for Imbalanced Problems	F1 Score	0.6003	0.5013	novel
Adversarial Classifier for Imbalanced Problems	Precision	0.8335	0.5371	novel
BAGAN: Data Augmentation with Balancing GAN	Balanced Accuracy	0.8068	0.5175	novel
BAGAN: Data Augmentation with Balancing GAN	SSIM	0.3226	0.6696	novel
Balancing thermal comfort datasets: We GAN, but should we?	Euclidean Distance	80.7967	41.8833	novel
Balancing thermal comfort datasets: We GAN, but should we?	F1 Score	0.6067	0.5247	novel
Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced le	Ranking	3.1548	4.0309	novel
DATA AUGMENTATION GENERATIVE ADVERSARIAL NETWORKS	Balanced Accuracy	0.6501	0.6937	baseline
Deep generative models to counter classimbalance: a model-metric mapping withpr	AUC	0.8460	0.8010	novel
Deep generative models to counter classimbalance: a model-metric mapping withpr	Balanced Accuracy	0.8448	0.8000	novel
Deep generative models to counter classimbalance: a model-metric mapping withpr	F1 Score	0.6600	0.5029	novel
Deep generative models to counter classimbalance: a model-metric mapping withpr	Geometric Mean	0.8173	0.7606	novel
Deep generative models to counter classimbalance: a model-metric mapping withpr	Precision	0.6510	0.5056	novel
Deep generative models to counter classimbalance: a model-metric mapping withpr	Recall	0.7118	0.7632	baseline
Effective data generation for imbalanced learning using conditional generative adve	Ranking	2.5100	4.2439	novel
EWGAN: Entropy-Based Wasserstein GAN for Imbalanced Learning	Balanced Accuracy	0.8354	0.7035	novel
FLOWGAN:Unbalanced network encrypted traffic identification method based on GA	Balanced Accuracy	0.9910	0.9394	novel
FLOWGAN:Unbalanced network encrypted traffic identification method based on GA	F1 Score	0.9910	0.9382	novel
FLOWGAN:Unbalanced network encrypted traffic identification method based on GA	Precision	0.9799	0.9455	novel
FLOWGAN:Unbalanced network encrypted traffic identification method based on GA	Recall	0.8995	0.9852	baseline
GAN-based imbalanced data intrusion detection system	Balanced Accuracy	0.9983	0.9930	novel
GAN-based imbalanced data intrusion detection system	F1 Score	0.9504	0.8782	novel
GAN-based imbalanced data intrusion detection system	Precision	0.9868	0.9758	novel
GAN-based imbalanced data intrusion detection system	Recall	0.9276	0.8404	novel
Generative adversarial fusion network for class imbalance credit scoring	AUC	0.7108	0.6491	novel
Generative adversarial fusion network for class imbalance credit scoring	Balanced Accuracy	0.8360	0.7713	novel
Generative adversarial fusion network for class imbalance credit scoring	F1 Score	0.5798	0.4439	novel
Generative Adversarial Minority Oversampling	Balanced Accuracy	0.7175	0.7673	baseline
Generative Adversarial Minority Oversampling	FID	3.4100	6.0600	novel
Generative Adversarial Minority Oversampling	Geometric Mean	0.6811	0.7240	baseline
Modeling Tabular Data using Conditional GAN	F1 Score	0.0895	-1.7357	novel
Modeling Tabular Data using Conditional GAN	Likelihood fitness metric	-6.9763	-10.9721	novel
PacketCGAN: Exploratory Study of Classimbalance for Encrypted Traffic Classificatio	Balanced Accuracy	0.9951	0.9805	novel
PacketCGAN: Exploratory Study of Classimbalance for Encrypted Traffic Classificatio	F1 Score	0.9947	0.9783	novel
PacketCGAN: Exploratory Study of Classimbalance for Encrypted Traffic Classificatio	Precision	0.9936	0.9792	novel
PacketCGAN: Exploratory Study of Classimbalance for Encrypted Traffic Classificatio	Recall	0.9958	0.9805	novel
Supervised Class Distribution Learning for GANs-based Imbalanced Classification	Balanced Accuracy	0.9880	0.7133	novel
Supervised Class Distribution Learning for GANs-based Imbalanced Classification	F1 Score	0.9520	0.6973	novel
Supervised Class Distribution Learning for GANs-based Imbalanced Classification	Geometric Mean	0.9740	0.7787	novel
Supervised Class Distribution Learning for GANs-based Imbalanced Classification	Precision	0.9540	0.6547	novel
Supervised Class Distribution Learning for GANs-based Imbalanced Classification	Recall	0.9640	0.5153	novel
Towards Class-Balancing Human Comfort Datasets with GANs	F1 Score	0.2773	0.3964	baseline
Unsupervised Anomaly Detection withGenerative Adversarial Networks to GuideMa	AUC	0.8900	0.7767	novel
Unsupervised Anomaly Detection withGenerative Adversarial Networks to GuideMa	Precision	0.8834	0.7986	novel
Unsupervised Anomaly Detection withGenerative Adversarial Networks to GuideMa	Recall	0.7278	0.6588	novel
Unsupervised Anomaly Detection withGenerative Adversarial Networks to GuideMa	Specificity	0.8298	0.8035	novel
Using generative adversarial networks for improving classification effectiveness in cr	recall	0.7194	0.7060	novel
Using generative adversarial networks for improving classification effectiveness in cr	Specificity	0.9999	1.0000	baseline

In 85% of experiments (as defined by unique paper-metric aggregations), the experimental method outperforms the baseline method

Therefore, domains where the total number of minority cases is lower (such as in financial fraud detection) may be more conducive to the use of AnoGAN architecture relative to CGAN architecture.

The preponderance of Wasserstein GANs and Wasserstein loss as a loss function in the domain of financial transactions relative to the domain of cybersecurity is related to the complexity of the respective datasets in these two domains. The implementation of Wasserstein GANs can provide greatest value in scenarios where the Generator can effectively fool the Discriminator without generating a synthetic example which is similar to the distribution in the actual training dataset. This is more likely to occur in datasets where the data is complex, the data types are many, and the various attributes display a latent interdependency. Thus, in the research works we reviewed, the use of Wasserstein loss was more prevalent in GAN methodologies applied to

Table 6 Average novel vs. baseline results, aggregated to the level of evaluation metric, with total number of occurrences of the metric in question, and the number of instances in which experimental methods were superior to baseline methods

Metric	Novel (avg)	Baseline (avg)	Perc. Improvement btwn. Novel and Baseline	Total Occurences	Occurences of Novel better than Baseline	Occurences of Baseline better than Novel
Balanced Accuracy	0.8853	0.8158	8.52%	12	10	2
F1 Score	0.6896	0.4532	52.17%	11	10	1
Precision	0.8975	0.7709	16.41%	7	7	0
Recall	0.8494	0.7785	9.11%	7	5	2
AUC	0.8613	0.8098	6.37%	5	4	1
Geometric Mean	0.8241	0.7544	9.24%	3	2	1
Ranking	2.8324	4.1374	46.07%	2	2	0
Specificity	0.9149	0.9018	1.46%	2	1	1
Euclidean Distance	80.7967	41.8833	92.91%	1	1	0
FID	3.4100	6.0600	77.71%	1	1	0
Likelihood fitness metric	-6.9763	-10.9721	-36.42%	1	1	0
SSIM	0.3226	0.6696	107.55%	1	1	0

financial transactions rather than cybersecurity problems, because the datasets in the former tend to be more heterogenous and less uniformly quantitative.

One fundamental question with which we may continue our analysis: how often did the researchers' experimental method "win"? That is, how often did their method best the baseline methods presented in the paper? We can answer this question in earnest using the information in Table 5. From here, we gauge that of the 53 unique combination aggregations of paper and evaluation metric, the experimental method outperforms the baseline methods in 45 of them (85%). Only [32] is responsible for multiple instances of baseline aggregations besting experimental aggregations (with balanced accuracy and geometric mean), and this research is a "groundwork" source related to images rather than tabular datasets. Two of these exceptions each feature recall and balanced accuracy as the evaluation metric, whereas AUC, specificity, F1 score, and geometric mean each account for one aggregated instance. Note that for Euclidean distance, a higher rather than lower distance is considered preferable. This is because for the paper in question [64], relating to human comfort datasets, Euclidean distance is used to quantify the difference between generated samples to measure the ability to avoid mode collapse; thus, samples with a high degree of variability (high Euclidean distance) are coveted. Only two research works ([39] and [46]) relevant to the use of GANs in class imbalance in tabular datasets reports the results at a more granular class level. Seven of the research works vary the factor of an imbalance ratio or number of minority samples generated, and report its effect on machine learning efficacy; this remains a potentially fruitful avenue for future work in the field.

We aggregate further to the level of performance metric in Table 6. With respect to frequency, we see that some balance-weighted form of accuracy and F1 score are the most popular evaluation metrics, with 12 and 11 occurrences, respectively. The structural similarity index in experimental methods improves most over baseline methods from a percentage perspective (107.55%), though the gains are limited to

Table 7 Average results by learner and evaluation metric for a subset of papers, experimental/novel methods vs. baseline methods

	AUC	Balanced Accuracy	F1 Score	Precision	Ranking	Recall	novel > baseline	baseline > novel
CNN							4	0
Novel		***0.6767***	***0.9947***	***0.9936***		***0.9958***		
Baseline		0.6187	0.9783	0.9792		0.9805		
DT							2	0
Novel			***0.9552***		***3.3370***			
Baseline			0.9482		4.0549			
GBC							1	0
Novel					***2.4850***			
Baseline					4.2097			
KNN							1	1
Novel			0.2600		***2.8140***			
Baseline			***0.2736***		4.1378			
LR							1	0
Novel					***3.6300***			
Baseline					3.9697			
MLP							3	1
Novel		***0.9937***	***0.9935***	***0.9890***		0.9637		
Baseline		0.9723	0.9703	0.9725		***0.9814***		
NB							0	1
Novel			0.1800					
Baseline			***0.4279***					
RF							2	0
Novel			***0.9543***		***2.5857***			
Baseline			0.9485		4.0488			
SAE							4	0
Novel		***0.9951***	***0.9947***	***0.9936***		***0.9958***		
Baseline		0.9805	0.9783	0.9792		0.9805		
SVM							1	2
Novel	0.7128		0.4275		***2.4633***			
Baseline	***0.9412***		***0.5231***		4.2556			
novel > baseline	0	3	5	3	6	2	19	
baseline > novel	1	0	3	0	0	1		5

Experimental methods have a "better" evaluation metric in 19 of 24 aggregation instances

its lone implementation in [80]. In contrast, F1 score improves by 52.17% through the use of experimental methods, and is used in 11 different papers. The gains for accuracy measures are positive but modest, at 8.52%. With seven occurrences apiece, precision and recall demonstrate gains over baseline methods of 16.41% and 9.11%, respectively. The former method boasts better experimental results in all seven of its occurrences, while F1 score outperforms baseline settings in all but one of its occurrences.

Table 7 provides the average results by evaluation metric (column) and learner (row) for experimental methods and baseline methods for the subset of papers explored in this survey. The balanced accuracy for experimental methods is higher across all three of the learners upon which it is evaluated than it is for baseline

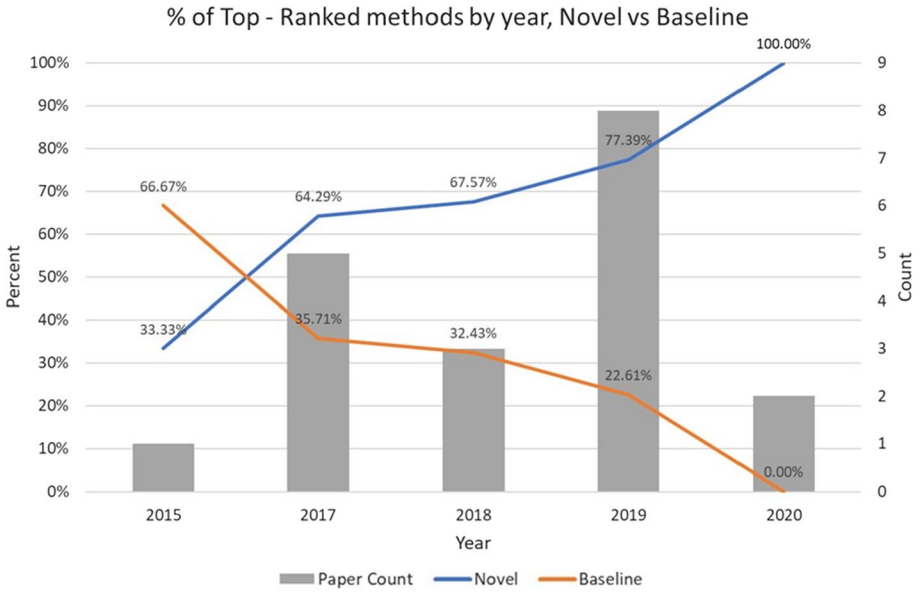

Fig. 3 Year by year aggregation of the percent share of methods in which experimental methods and baseline methods were the top-ranked approach, 2015–2020 (excluding 2016). The paper count for the respective year is also given in the corresponding bar chart

methods. Likewise for precision as an evaluation metric and ranking, in which all experimental methods had better average rankings across all six of the learners with which they were affiliated. The use of F1 score was more mixed (five in favor of experimental methods, three in favor of baseline methods) though still advantageous to experimental methods, while the lone usage of the AUC in this subset yielded a higher baseline evaluation (with an SVM learner). Similarly, the CNN (4), SAE (4), Decision Tree (2) and Random Forest (2) learners had better-rated experimental methods across all of their (parenthetical) respective learners. For these learners, the machine learning efficacy of methods with experimental data sampling procedures (that is, the implementation of GANs specific to each paper) frequently bested the machine learning efficacy of methods with baseline data sampling procedures (ROS, SMOTE, more rudimentary GANs, etc.).

The general thrust of related research suggests GANs and their increasingly advanced methods are accomplishing classification feats that their "traditional" counterparts are unable to match. Figure 3 provides a yearly aggregation of the percent share of methods in which experimental methods and baseline methods were the top-ranked approach. The paper count for each year is also given, to provide quantitative context. For papers published in 2015, 66.67% of the experiments had a baseline method as the top-ranked method, compared to 33.33% with an experimental method ranked most highly. In this section of analysis, no papers published in 2016 were explored, but the superiority of experimental methods re-asserts itself in 2017, with these methods now constituting 64.29% of top-ranked methods. This upward trend continues monotonically until the most recently analyzed year in 2020, where all of the experiments in both of the papers studied boasted novel methods as the top-ranked method.

Table 8 Citation count for Top 20 cited papers used in this survey, as ranked by frequency

Rank	Paper	Count
1	Generative adversarial networks	31
2	Unsupervised representation learning with deep convolutional generative adversarial networks	16
3	Smote: Synthetic minority oversampling technique	15
4	Conditional generative adversarial nets	11
5	Deep Residual Learning for ImageRecognition	9
6	Deep generative image models using a laplacian pyramid of adversarial networks	8
6	Bagan: Data augmentation with balancing gan	8
6	Improved techniques for training GANs	8
9	Learning deep representation for imbalanced classification	7
9	Learning multiple layers of features from tiny images	7
9	Unpaired imagetoimage translation using cycleconsistent adversarial networks	7
9	Conditional image synthesis with auxiliary classifier GANs.	7
9	BorderlineSMOTE: A new oversampling method in imbalanced data sets learning	7
9	ADASYN: Adaptive synthetic sampling approach for imbalanced learning	7
15	AutoEncoding VariationalBayes	6
15	Data augmentation generative adversarial networks	6
15	Effective data generation for imbalanced learning using Conditional Generative Adversarial Networks	6
15	ImageNet classification with deep convolutionalneural networks	6
15	Learning from imbalanced data	6
20	Wasserstein gan	5

Citation network analysis

To address prominent ideas and patterns in the body of research abreast the intersection of GANs and class imbalance, we create and analyze a directed citation network created by a subset of the papers cited in this survey. This initially includes an unrestricted citation count of all papers cited in this survey, and then node/entity-level metrics of the network formed by the aforementioned subset. We also present the graphical structure created by this citation network.

As a basic measure we can examine the papers with the highest citation count. Unsurprisingly, this analytical lens, the output of which is given in Table 8, shows the research of [9] to be the seminal work in the field. Moving further down the list helps trace the evolution and best practices in the discipline. The work of [81], the second most cited paper, serves as a direct descendant of the bedrock GAN approach, representing its first noteworthy application in an end-to-end classification scheme with requisite architecture modifications. Our third most cited paper, [19] is the most frequently-exploited non-generative baseline method, applicable to both visual and tabular problems. Likewise, [20] enhances the SMOTE algorithm and is cited 7 times, while ADASYN [82] offers another synthetic non-adversarial data generation method and is also cited 7 times. Subsequently, [83], another immediate successor to the original GAN methodology, is cited 11 times, with an important modification allowing for a mechanism of control over which class of samples the Generator may generate. Similarly, [31]—with 6 citations—stands on the shoulders of CGAN and ACGAN, bolstered by autoencoding techniques to avoid mode collapse and by a dual-label Discriminator which outputs either "fake" or the predicted class label. [56] offers another common modification on

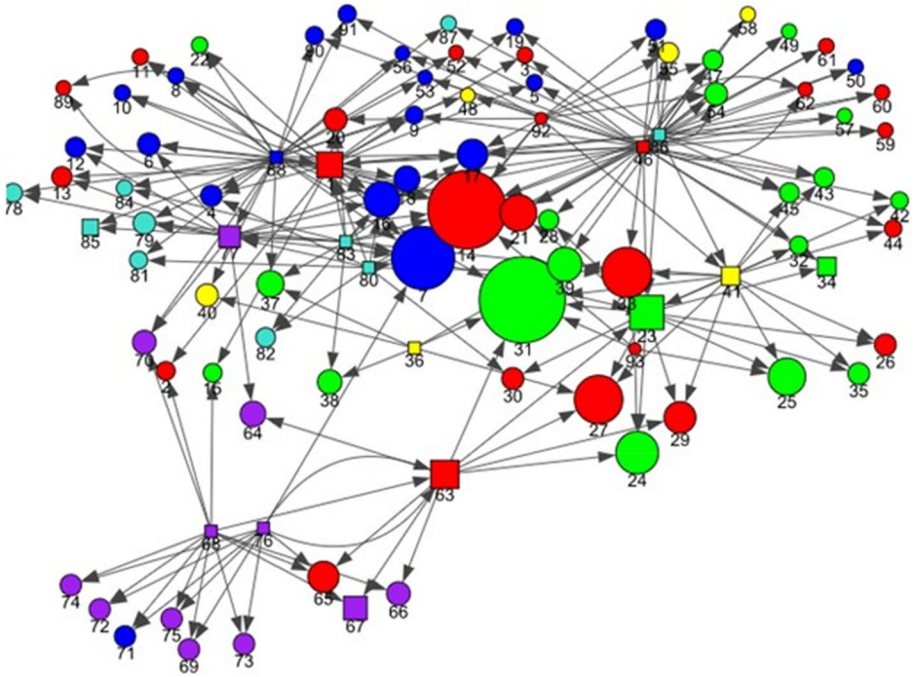

Fig. 4 Directed network (digraph) of a subset of research papers explored in this survey, and the corresponding works these papers cite. Node size is related to calculated Pagerank values

orthodox GAN architecture, utilizing the Wasserstein loss during Generator and Discriminator training.

In Fig. 4, we use a subset of twenty papers presented in this work to create a research citation network. In general, the papers we elected for this subset were those more directly pertinent to the question of GANs and class imbalance in tabular data. From the compiled bibliographies of these selected papers, we created a directed graph (digraph) where the paper in question served as the head, and the cited paper served as the tail. Selected papers—that is, those addressed in this survey—are represented by square or rectangular nodes, while all other papers are represented by circular nodes. This edge list was cobbled together and fed into Python in the Jupyter IDE, where the "igraph" package was used for analysis and presentation. For presentational clarity, blue nodes are those papers which correspond to baseline methods, yellow nodes are those papers which correspond to healthcare, purple nodes are those papers which correspond to cybersecurity, turquoise nodes are those which correspond to financial fraud, and red nodes are those which correspond to all other papers. This ultimately resulted in a digraph with 93 nodes (research papers), given in Fig. 4; the corresponding ID assigned in Tables 10 and 11 in Appendix is used interchangeably for expositional brevity.

The sizes of the nodes in the digraph in Fig. 4 are positively related with the Pagerank value of the nodes, given in Table 9 with the corresponding node ID.[3] Here, we see the most influential papers are [9, 19], and [81], dealing with the introduction of GANs, SMOTE, and GANs in fully-pipelined classification schemes. This largely confirms the

[3] The node IDs are given as labels in the digraph, directly below the node in question. A list of these IDs is reproduced in the Appendix.

Table 9 Pagerank, degree scores, and betweenness scores for the various nodes (node IDs) in the digraph

ID	Pagerank	In Degree	Out Degree	Total Degree	Betweenness	ID	Pagerank	In Degree	Out Degree	Total Degree	Betweenness	ID	Pagerank	In Degree	Out Degree	Total Degree	Betweenness
1	0.01137	5	21	26	63	32	0.01039	2	0	2	0	63	0.01159	3	9	12	32
2	0.01045	2	0	2	0	33	0.01339	5	0	5	0	64	0.01131	2	0	2	0
3	0.01006	2	0	2	0	34	0.01039	2	0	2	0	65	0.01188	3	0	3	0
4	0.01081	3	0	3	0	35	0.01086	2	0	2	0	66	0.01109	2	0	2	0
5	0.01006	2	0	2	0	36	0.00938	0	6	6	0	67	0.01109	2	0	2	0
6	0.01093	3	0	3	0	37	0.01143	3	0	3	0	68	0.00938	0	13	13	0
7	0.01419	9	0	9	0	38	0.01118	2	0	2	0	69	0.01079	2	0	2	0
8	0.01009	2	0	2	0	39	0.01219	4	0	4	0	70	0.01108	3	0	3	0
9	0.01030	3	0	3	0	40	0.01096	2	0	2	0	71	0.01079	2	0	2	0
10	0.01009	2	0	2	0	41	0.01044	2	13	15	24.5	72	0.01079	2	0	2	0
11	0.01009	2	0	2	0	42	0.01028	2	0	2	0	73	0.01079	2	0	2	0
12	0.01068	2	0	2	0	43	0.01061	3	0	3	0	74	0.01079	2	0	2	0
13	0.01068	2	0	2	0	44	0.01028	2	0	2	0	75	0.01079	2	0	2	0
14	0.01499	9	0	9	0	45	0.01061	3	0	3	0	76	0.00938	0	10	10	0
15	0.01045	2	0	2	0	46	0.00938	0	37	37	0	77	0.01082	3	11	14	13
16	0.01233	6	0	6	0	47	0.01065	3	0	3	0	78	0.01047	2	0	2	0
17	0.01183	6	0	6	0	48	0.00984	2	0	2	0	79	0.01093	3	0	3	0
18	0.01136	4	0	4	0	49	0.00993	2	0	2	0	80	0.00938	0	11	11	0
19	0.01017	2	0	2	0	50	0.00993	2	0	2	0	81	0.01035	2	0	2	0
20	0.01103	3	0	3	0	51	0.01065	3	0	3	0	82	0.01057	2	0	2	0
21	0.01238	6	0	6	0	52	0.00984	2	0	2	0	83	0.00938	0	17	17	0
22	0.01009	2	0	2	0	53	0.00984	2	0	2	0	84	0.01010	2	0	2	0
23	0.01217	5	13	18	66.5	54	0.01098	4	0	4	0	85	0.01010	2	0	2	0
24	0.01286	3	0	3	0	55	0.01065	3	0	3	0	86	0.00938	0	24	24	0
25	0.01245	3	0	3	0	56	0.00984	2	0	2	0	87	0.00996	2	0	2	0
26	0.01086	2	0	2	0	57	0.00993	2	0	2	0	88	0.00938	0	32	32	0
27	0.01328	4	0	4	0	58	0.00993	2	0	2	0	89	0.00988	2	0	2	0
28	0.01064	3	0	3	0	59	0.00993	2	0	2	0	90	0.01035	2	0	2	0
29	0.01195	3	0	3	0	60	0.00993	2	0	2	0	91	0.01035	2	0	2	0
30	0.01090	2	0	2	0	61	0.00993	2	0	2	0	92	0.00938	0	11	11	0
31	0.01534	7	0	7	0	62	0.00981	2	0	2	0	93	0.00938	0	5	5	0

observations made earlier in this section regarding the bedrock importance of these papers to the treatment of GANs in tabular class imbalanced settings. These papers also have the highest in-degree of papers which cite them. Borderline-SMOTE and ADAYSN are both frequently cited as baseline methods, and consequently both [20] and [14] possess a high in-degree count. The use of conditional GANs for data generation based on class labels is addressed in [52, 79], and [84] (all with five or six in-degrees). With five in-network citations, [85] offers advice vis-à-vis best practices for hyper-tuning GANs at time of publication, though this work is mainly done with computer vision tasks in mind.

Out-degree is also reported in Table 9, though a discussion on this metric is trivial given such linkages are limited toward the handful of papers we designated as seed papers for this network analysis. Though betweenness measures suffer from a similar limitation (i.e., a node can only have a non-zero betweenness measure if it is one we have designated as a seed paper), there is an extended value in highlighting papers with positive betweenness measures, as these serve to "connect" two or more separate works which otherwise share no citation kinship.[4] [36] is a prominent work which uses GANs to address the imbalance question in network traffic classification (betweenness score of 32), while [49] does the same for fraud detection (13). With a betweenness score of 24.5, [31] is often used in tabular GAN-based schemes as a modified architecture of CGAN.

Discussion

The use of GANs as potential alleviants to class imbalance has garnered sizable research interest, but this attention has primarily been reserved for computer vision tasks. As such, this specific implementation has been muted on tabular datasets. The recipient domains of generative data augmentation methods thus far have primarily been cybersecurity and financial fraud detection. Datasets pertaining to IoT applications have also leveraged GANs to enrich training data, but in many scenarios, this is not done explicitly in the context of class imbalance, but rather to create synthetic data to replace original data which is presumed inaccurate [86].

[4] Graphically speaking, the shortest path between two other nodes passes through the node in question.

There are number of unique challenges in applying GANs to tabular data, challenges not previously addressed by counterpart problems in visual domains. The heterogeneity of data types poses one of the most common issues; in a given dataset, columns may be numerically discrete, numerically continuous, binary, or categorical. Though a common treatment to numeric columns is to apply a min–max transformation to these attributes, continuous values in tabular datasets are often non-Gaussian, which Xu et al. [71] point out can evoke the dilemma of vanishing gradients. Another complication with numeric attributes can come in bi-modal distributions, which are fairly common structural representations of income or price values. Likewise for categorical explanatory columns, class imbalance can lead to severe mode collapse at the expense of the minority class, and to the detriment of an untrained Discriminator.

With respect to categorical variables, the Gumbel softmax function has arisen as a popular quantification mechanism for consumption by GANs. Ultimately, for a given categorical variable, the Gumbel softmax function generates one-hot encodings with added noise to avoid being fully differentiable by the Discriminator during training. Of particular importance in the function is the temperature- τ parameter, which controls the diversity of the output, and could prove an interesting factor to manipulate in future research. For numerical attributes of interval or ratio scales, min–max scaling is often applied, with small additions of Gaussian noise to allay any potential identity relationship between columns that might otherwise be learned and exploited by the Generator.

In contrast to research tasks involving computer vision, the lack of local structure amongst the explanatory columns in tabular data has made it difficult to craft one-size-fits-many convolutional pipelines of Generators and Discriminators that can be generalized to different domains, or even different research problems within the same domain. Indeed, even for the research papers discussed in this survey that experienced success with the use of GANs in tabular problems, the preprocessing methods utilized to make datasets consumable by GAN components cannot be neatly categorized into a series of generalizable steps; that is, for hypothetical Research Group 1, Column A is an integer type requiring whole numbers and generated instances must reflect this, Column B is a categorical string type with Bayesian dependence on Column A that must shine through in generated instances to fool the Discriminator, Column C is a float type with bi-modal distribution, etc. Meanwhile, for hypothetical Research Group 2, the preprocessing steps taken to homogenize its dataset were entirely different.

These problem-specific, bespoke adjustments made by researchers in the course of their respective research processes are not at all a criticism of these processes. Rather, they speak to the need to craft more generalizable data preprocessing methodologies which can ingest tabular datasets of mixed data types, and output a tabular dataset which can be readily consumed by a GAN. Alternatively, this wide disparity in methodological approaches points to the need for tabular GAN architectures which can ingest data in relatively unprocessed states, and perform all necessary transformations "under the hood" without the need for such diverse, personalized data preprocessing machinations by researchers. To the best of our knowledge, the only such methodologies with widespread adoption across different domains are Tabular-GAN, TabNet, and table-GAN [87], though even they leave something to be desired with respect to the need for data preprocessing.

Though this survey has demonstrated the ability of GANs to correct imbalance in tabular settings, the use of GANs is not appropriate, or at the very least not necessary, in all research settings. GANs and other deep learning approaches require extensive optimization of hyperparameters for proper implementation, and the proper hardware and software provisions to accomplish this optimization. In environments with limited computing resources, the use of traditional synthetic methods such as SMOTE paired with gradient boosting and random forests for classification may be a more feasible alternative, though researchers such as Okerinde et al. [88] are experimenting with incorporating GANs into transfer learning environments through the use of encoder modules. Additionally, even where the resources to support GAN pipelines are in place, and where GANs outperform traditional resampling methods, the lift provided by GAN methods may not justify the added procedural complexity. Oftentimes, small preprocessing tweaks to the dataset to make the data less complex can appreciably boost the performance of traditional methods, and thereby decrease the marginal benefit of using GANs to correct class imbalance. Quintana and Miller [64] make a similar observation, noting that consolidating their categorical variable of interest from seven classes to three classes significantly reduced the machine learning efficacy advantage of their novel ComfortGAN over simpler baseline methods.

Experimentally, one factor of great interest which only a few current publications consider is the optimal level of class imbalance and data generation, given an original level of class imbalance. That is, how many samples of a minority class should a GAN generate to render maximum machine learning efficacy? Alternatively, in terms of a percent or ratio, by what factor can the representation of minority instances increase before one observes declines in machine learning efficacy evaluation metrics? Khoshgoftaar et al. [89] have previously considered the question for traditional sampling methods, but few successor studies have posed the same interrogative with respect to GANs and tabular data.

Fiore et al. [49] and Lei et al. [51] are the only two tabular methodologies in this survey to consider the level of class imbalance (and therefore the level of data generation performed by the GAN) as an experimental factor. Though the former does not emphasize machine learning efficacy with respect to variation of this factor, an increased number of generated minority instances generally increases precision and F-measure; however, the increase is concave, with the most balanced performance coming with fourfold to eightfold generated increases of minority instances. Likewise, Lei et al. [51] find that AUC, balanced accuracy, and F1 score most improve when the injection of synthetic minority instances increases minority representation by a factor of 1.5 or 2, after which there is a notable decline in each of these evaluation metrics. Future studies should focus on this parameter as one of primary interest, and whether this range—minority multiplication by a factor of anywhere between 2 and 8—holds across other domains, such as cybersecurity. Similarly, future research endeavors may ask the related question of whether GANs will be able to offer any performance gain in cases/classes of extreme rarity (say, fewer than five instances).

Because of the similarity in objective between anomaly detection and one-class classification, an open research question is how GAN-infused datasets can be leveraged to improve reconstruction methods, boundary methods, and density estimation methods

in one-class classification. In a sense, the AnoGAN family of GAN methodologies relies on a similar concept of "pre-filtering" the dataset for the desired class, and training neural networks (the Generator and the Discriminator) on identifying the structural patterns in this class. However, that particular class of implementation speaks only to the generative component of the GAN pipeline. Though various research works in computer vision have taken up the research question of whether synthetically augmented datasets can improve one-class classification systems, at present the literature is sparse for the same exercise with tabular data.

Another element largely missing from the literature is a method to gauge the legitimacy of the data generated by GANs independent of the machine learning efficacy results. That is, the gap remains for the development of tabular analogs to metrics such as the Fréchet inception distance [90] to judge the quality of synthetic data. For example, Sajjadi et al. [91] examined a systematic way to evaluate the instances produced by GANs in regimes of class imbalance, suggesting a system whereby, given an actual distribution P and a synthesized distribution Q, recall measures the proportion of P covered by Q, while precision measures the quality of the samples in Q. Moreover, [91] posit that the use of these twin metrics can protect against mode collapse without the need for a manual inspection. However, this lens by which to evaluate data quality has yet to be applied to tabular datasets, and many other similar methodologies can help ensure the data generated by GANs represents an appropriate integration into the original dataset.

Conclusion

This paper presented a detailed tour through the landscape of state-of-the-art methods by which GANs help correct class imbalance in tabular datasets. After a preparatory retracing of GANs and class imbalance as individual topics, we then turned our attention to more tabular concerns, focusing on contributions in cybersecurity and financial transactions. We also looked at an assemblage of different topics, some in different domains, others procedural. After this, we performed a meta-analysis of the papers surveyed herein, noting the preponderance of evaluation metrics, datasets, baseline methods, learners, and the like, as well as the frequency with which experimental methods bested baseline methods. Finally, we undertook a network analysis of the citation network created by a subset of the papers utilized in this survey, noting the research works and associated concepts which are most influential in the field. In general, we found GANs have enjoyed notable success in restoring balance to these datasets, though the experiments in question often deprived us an experimental viewing of GANs in multiclass settings, lacked procedural similarity to compare across architectures, and generally ignored the question of the optimal amount level or percent of injection of GAN-generated instances for machine learning efficacy. These and many other avenues constitute potential future work for the implementation of GANs on imbalanced tabular data.

Appendix

See Tables 10, 11.

Table 10 ID labels used in digraph above, with corresponding paper names, Part I

Paper ID	Paper Name
1	Effective data generation for imbalanced learning using conditional generative adversarial networks
2	TensorFlow: Largescale machine learning on heterogeneous systems
3	Mwmote– majority weighted minority oversampling technique for in balanced data set learning
4	A study of the behavior of several methods for balancing machine learning training data
5	Safelevelsmote: Safelevelsynthetic minority over sampling technique for handling the class imbalanced problem
6	DBSMOTE: Densitybased synthetic minority oversampling technique
7	SMOTE: Synthetic minority oversampling technique
8	Data mining for imbalanced datasets: An overview
9	Smoteboost: Improving prediction of the minority class in boosting
10	Start globally optimize locally predict globally: Improving performance on imbalanced data
11	Selforganizing map oversampling (SOMO) for imbalanced data set learning
12	A review on ensembles for the class imbalance problem: bagging– boosting– and hybrid–based approaches
13	Deep sparse rectifier neural networks
14	Generative Adversarial Networks
15	Learning from imbalanced data sets with boosting and data generation: The DataBoost IM approach
16	BorderlineSMOTE: A new oversampling method in imbalanced data sets learning
17	Adasyn: Adaptive synthetic sampling approach for imbalanced learning
18	Learning from imbalanced data
19	Adam: A method for stochastic optimization
20	Imbalancedlearn: A python toolbox to tackle the curse of imbalanced datasets in machine learning
21	Conditional Generative Adversarial Nets
22	Adaptive semiunsupervised weighted oversampling (ASUWO) for imbalanced datasets
23	Conditional Image Synthesis with Auxiliary Classifier GANs
24	Deep generative image models using a laplacian pyramid of adversarial networks
25	AdversarialFeature Learning
26	AdversariallyLearned Inference
27	Generative adversarial nets
28	AutoEncoding VariationalBayes
29	Synthesizing the preferred inputs for neurons in neural networks via deep generator networks
30	SemiSupervised Learning with Generative Adversarial Networks
31	Unsupervised representation learning with deep convolutional generative adversarial networks
32	Stochastic Backpropagation and Approximate Inference in DeepGenerative Models
33	Improved techniques for training GANs
34	Unsupervised and semisupervised learn ing with categorical generative adversarial networks
35	Image quality assessment: from error visibility to structural similarity
36	Data Augmentation Generative Adversarial Networks
37	Wasserstein GAN
38	Improved Training of Wasserstein GANs
39	Deep Residual Learning for ImageRecognition
40	Imagenet classification with deep convolutional neural networks
41	BAGAN: Data Augmentation with Balancing GAN
42	VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning
43	Gradientbased learning applied to document recognition
44	Deligan: Generative adversarial networks for diverse and limited data
45	Learning multiple layers of features from tiny images
46	Generative Adversarial Minority Oversampling
47	Deep oversampling framework for classifying imbalanced data

Table 11 ID labels used in digraph above, with corresponding paper names, Part II

Paper ID	Paper Name
48	Began: boundary equilibrium generative adversarial networks
49	A survey of predictive modeling on imbalanced domains
50	Costaware pretraining for multiclass costsensitive deep learning
51	Imbalanced deep learning by minority class incremental rectification
52	SMOTE for learning from imbalanced data: Progress and challenges marking the 15year anniversary
53	Gans trained by a two timescale update rule converge to a local nash equilibrium
54	Learning deep representation for imbalanced classification
55	Learning from imbalanced data: open challenges and future directions
56	Least squares generative adversarial networks
57	Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance
58	Reading digits in natural images with unsupervised feature learning
59	A classificationbased study of covariate shift in gan distributions
60	Learning to model the tail
61	Fashionmnist: a novel image dataset for benchmarking machine learning algorithms
62	Holisticallynested edge detection
63	A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification
64	Deep learning
65	Learning from imbalanced data for encrypted traffic identification problem
66	Service Name and Transport Protocol Port Number Registry
67	Towards automated application signature generation for traffic identification
68	FLOWGAN:Unbalanced Network Encrypted Traffic Identification Method Based on GAN
69	Mobile encrypted traffic classification using deep learning
70	The class imbalance problem: a systematic study
71	Datanet: Deep learning based encrypted network traffic classification in sdn home gateway
72	Endtoend encrypted traffic classification with onedimensional convolution neural networks
73	Network traffic classifier with convolutional and recurrent neural networks for internet of things
74	A hierarchical approach to encrypted data packet classification in smart home gateways
75	Characterization of encrypted and vpn traffic using timerelated features
76	PacketCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN
77	Using generative adversarial networks for improving classification effectiveness in credit card fraud detection
78	Calibrating probability with undersampling for unbalanced classification
79	An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics
80	Generative adversarial fusion network for class imbalance credit scoring
81	Balancing training data for automated annotation of keywords: a case study
82	Benchmarking stateoftheart classification algorithms for credit scoring: an update of research
83	Conditional Wasserstein GANbased oversampling of tabular data for imbalanced learning
84	A StyleBased Generator Architecture for Generative Adversarial Networks
85	Modeling Tabular data using Conditional GAN
86	Supervised Class Distribution Learning for GANsBased Imbalanced Classification
87	Learning from classimbalanced data: Review of methods and applications
88	Deep Generative Models to Counter Class Imbalance: A ModelMetric Mapping With Proportion Calibration Methodology
89	Improving imbalanced learning through a heuristic oversampling method based on kmeans and SMOTE
90	Borderline oversampling for imbalanced data classification
91	An instance level analysis of data complexity
92	Adversarial Classifier for Imbalanced Problems
93	Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery

Abbreviations

GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
CGAN	Conditional Generative Adversarial Network
SMOTE	Synthetic Minority Oversampling Technique
BAGAN	Balanced Generative Adversarial Network
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
ROS	Random Oversampling
RUS	Random Undersampling
CBOS	Cluster-Based Oversampling
ADASYN	Adaptive Synthetic
AC-GAN	Auxiliary Classifier Generative Adversarial Network
SSIM	Structural Image Similarity
GAMO	Generative Adversarial Minority Oversampling
ACSA	Average Class Specific Accuracy
GM	Geometric Mean
NIMS	Network Information Management and Security
SSH	Secure Shell
AUC	Area Undeath Receiver-Operator Curve
CICIDS	Intrusion Detection Evaluation Dataset
RF	Random Forest
ISCX	Intrusion Detection Evaluation
IoT	Internet of Things
M2M	Machine to Machine
WGAN	Wasserstein Generative Adversarial Network
PRC	Precision—Recall Curve
PGAN	Polarity Generative Adversarial Network
VAE	Variational Autoencoder
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty
cWGAN	Conditional Wasserstein Generative Adversarial Network
VGM	Variational Gaussian Mixture Model
MNIST	Modified National Institute of Standards and Technology
EWGAN	Entropy-based Wasserstein Generative Adversarial Network
ARIC	Adversarially Re-weighting for Imbalanced Classification
MMM	Model-Metric Mapper methodology
CIFAR	Canadian Institute for Advanced Research
KNN	K-Nearest Neighbors
DT	Decision Tree
LR	Logistic Regression
GBC	Gradient Boosting Classifier with Decision Trees as Base Learners
SAE	Sparse Autoencoder
NB	Naive Bayes
FID	Frechet Inception Distance
SVHN	Street View House Numbers
RBM	Restricted Boltzmann Machine
AnoGAN	Anomaly Generative Adversarial Network
IDE	Integrated Development Environment
UCI	University of California at Irvine
DAGAN	Data Augmentation Generative Adversarial Network
SGAN	Semi-supervised Generative Adversarial Network
ReLU	Rectified Linear Unit
SMOTE-ENC	Synthetic Minority Oversampling Technique Encoded Nominal and Continuous
CTGAN	Conditional Tabular Generative Adversarial Network

Acknowledgements
We would like to thank the various reviewers in the Data Mining and Machine Learning Laboratory at Florida Atlantic University, Boca Raton, FL 33431, particularly Justin Johnson, Dr. Rich Bauder, Dr. Karl Weiss, and Dr. Naeem Seliya.

Author contributions
RSC surveyed and summarized the various works relating to the topic. RSC also compiled and tabularized the metadata related to the inspected works and generated the output for the various tables and figures used in the presented analyses. TMK guided the direction of the research and helped to finalize the work. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 9 December 2021 Accepted: 7 June 2022
Published online: 22 August 2022

References

1. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis. 2002;6(5):429–49.
2. Japkowicz N. The Class Imbalance Problem: Significance and Strategies. In: Proc. of the Int'l Conf. on Artificial Intelligence, 2000.
3. Liu X-Y, Zhou Z-H, Wu J. Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics. Part B (Cybernetics). 2009;39(2):539–50.
4. Guo X, Yin Y, Dong C, Yang G, Guangtong Z. On the Class Imbalance Problem. In: 2008 Fourth International Conference on Natural Computation, 2008.
5. Anand R, Mehrotra KG, Mohan CK, Ranka S. An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets Rangachari h a n. In: IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 4, no. 6, 1993.
6. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.
7. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249–59.
8. Ren M, Zeng W, Yang B, Urtasun R. Learning to Reweight Examples for Robust Deep Learning. In: Proceedings of the 35th International Conference on Machine Learning, p. 4334–4343, 2018.
9. Goodfellow IJ, Pouget-Abadie J, Mizra M, Xu B, Warde-Farley D, Ozair S, Courville and Y. Bengio, "Generative Adversarial Networks. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), p. 2672–2680, 2014.
10. Scott M, Plested J. GAN-SMOTE: A Generative Adversarial Network approach to Synthetic Minority Oversampling for One-Hot Encoded Data. In: ICONIP2019 Proceedings, 2019.
11. Haldar M, Abdool M, Ramanathan P, Xu T, Yang S, Duan H, Zhang Q, Barrow-Williams N, Turnbull BC, Collin BM, Legrand T. Applying Deep Learning To Airbnb Search," in KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019.
12. Arik S, Pfister T. TabNet: Attentive Interpretable Tabular Learning. In: Association for the Advancement of Artifical Intelligence; 2020.
13. Popov S, Morozov S, Babenko A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. In: International Conference on Learning Representations; 2019.
14. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
15. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progr Artif Intell. 2016;5:221–32.
16. Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179 - 186, 1997.
17. Chawla NV, Japkovicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. In: ACM SIGKDD Explorations Newsletter; vol. 6, no. 1, 2004.
18. Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. In: ICML '07: Proceedings of the 24th international conference on Machine learning, p. 935–942, 2007.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:331–57.
20. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Lecture Notes in Computer Science; 2005. p. 878–87.
21. Jo T, Japkovicz N. Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsl. 2004;6(1):40–9.
22. Wilson DL. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Trans Syst Man Cybern. 1972;2(3):408–21.
23. Tomek I. Two Modifications of CNN. IEEE Trans Syst Man Cybern. 1971;6(11):769–72.
24. Tsai C-F, Lin W-C, Ke S-W. Big data mining with parallel computing: a comparison of distributed and MapReduce methodologies. J Syst Softw. 2016;122:83–92.
25. Yin L, Ge Y, Xiao K, Wang X, Quan X. Feature selection for high-dimensional imbalanced data. Neurocomputing. 2013;105:3–11.
26. Miller Al. Ian Goodfellow's Generative Adversarial Networks: AI Learns to Imagine. Cambridge: MIT Press; 2019.
27. Wang Z, She Q, Ward TE. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," ACM Computing Survey, 2020.
28. Sampath V, Maurtua I, Aguilar Martín JJ, Gutierrez A. A survey on generative adversarial networks for imbalance problems in computer vision tasks. In: J Big Data; 2021.
29. Odena A, Olah C, Shlens J. Conditional Image Synthesis with Auxiliary Classifier GANs. In: Proceedings of the 34 th International Conference on Machine Learning; 2017.

30. Antoniou A, Storkey A, Edwards H. Data Augmentation Generative Adversarial Networks. In: International Conference on Learning Representations; 2018.
31. Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C. BAGAN: Data Augmentation with Balancing GAN. ArXiv, abs/1803.09655; 2018.
32. Mullick SS, Datta S, Das S. Generative Adversarial Minority Oversampling. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019. p. 1695–704.
33. Ando S, Huang CY. Deep Over-sampling Framework for Classifying Imbalanced Data. In: Lecture Notes in Computer Science , vol. 40534; 2017.
34. Cieslak DA, Chawla NV, Striegel A. Combating Imbalance in Network Intrusion Datasets. In: 2006 IEEE International Conference on Granular Computing 2006. p. 732–7.
35. Khoshgoftaar TM, Leevy JL. A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data. J Big Data; 2020.
36. Vu L, Bui CT, Nguyen U. A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification. In: SolCT; 2017.
37. Alshammari R, Zincir-Heywood AN. Can encrypted traffic be identified without port numbers, IP addresses and payload inspection? Comput Netw. 2010;55(6):1326–50.
38. More A. Survey of resampling techniques for improving classification performance in unbalanced datasets. In: Computing Research Repository, vol. abs/1608.06048, 2016.
39. Lee J, Park K. GAN-based imbalanced data intrusion detection system. Pers Ubiquit Comput. 2019;25:121–8.
40. Sharafaldin I, Lashkari AH, Ghorbani AA. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In: 4th International Conference on Information Systems Security and Privacy (ICISSP); 2018.
41. Wang Z, Wang P, Zhou X, Li S, Zhang M. FLOWGAN:Unbalanced network encrypted traffic identification method based on GAN. In: 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom); 2019.
42. Draper-Gil G, Lashkari AH, Mamun MSI, Ghorbani AA. Characterization of Encrypted and VPN Traffic using Time-relatedFeatures. In: International Conference on Information Systems Security and Privacy (ICISSP 2016). p. 407–14.
43. Wang P, Li S, Ye F, Wang Z, Zhang M. PacketCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN. In: ICC 2020 - 2020 IEEE International Conference on Communications (ICC); 2020. p. 1–7.
44. Wang W, Zhu M, Zeng X, Ye X, Sheng Y. Malware traffic classification using convolutional neural network for representation learning. In: International Conference on Information Networking; 2017.
45. Yilmaz I, Masum R, Siraj A. Addressing Imbalanced Data Problem with Generative Adversarial Network For Intrusion Detection. In: 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI); 2020.
46. Macia Ferndandez G, Camacho J, Magan-Carrion R, Garcia-Teodoro P, Theron R. Ugr'16: a new dataset for the evaluation of cyclostationarity-based network IDSs. In: Computers & Security; 2017.
47. Belenko V, Chernenko V, Kalinin M, Krundyshev V. Evaluation Of GAN Applicability for Intrusion Detection in Self-Organizing Networks of Cyber Physical Systems. In: 2018 International Russian Automation Conference (RusAutoCon); 2018.
48. Jegadeesan K, Ayothi S. An Empirical Study of Methods, Metrics and Evaluation of Data Mining Techniques in Credit Card Fraudulence Detection. J Adv Res Dynam Control Syst. 2020;12:7.
49. Fiore, U, De Santis A, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. In: Information sciences; 2019. p. 448–55.
50. Dal Pozzolo A, Caelen O, Bontempi G. Calibrating Probability with Undersamplingfor Unbalanced Classification. In: IEEE Symposium Series on Computational Intelligence, 2015.
51. Lei K, Xie Y, Zhong S, Dai J, Yang M, Shen Y. Generative adversarial fusion network for class imbalance credit scoring. Neural Comput Appl. 2020;32:8451–62.
52. Odena A. Semi-Supervised Learning with Generative Adversarial Networks. In: Data Efficient Machine Learning workshop at ICML 2016, 2016.
53. Yeh I-C, Lien C-H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Sys Appl. 2009;36(2):2473–80.
54. Engelmann J, Lessmann S. Conditional Wasserstein GAN-based oversampling of tabular data for Imbalanced Learning. In: Expert Systems With Applications, 2021.
55. Quintana M, Miller C. Towards Class-Balancing Human Comfort Datasets with GANs. In: The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19); 2019.
56. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. In: International conference on machine learning. PMLR; 2017. p. 214–23.
57. Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. In: International Conference on Learning Representations; 2017.
58. Mottini A, Lheritier A, Acuna-Agost R. Airline Passenger Name Record Generation using Generative Adversarial Networks. In: ICML 2018 - workshop on Theoretical Foundations and Applications of Deep Generative Models; 2018.
59. Lopez V, Fernandez A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci. 2013;250:113–41.
60. Wang C, Yu Z, Zheng H, Wang N, Zheng B. CGAN-PLANKTON: Towards Large-Scale Imbalanced Class Generation and Fine-Grained Classification. In: 2017 IEEE International Conference on Image Processing (ICIP); 2017. p. 855–9.
61. Orenstein ECC, Beijbom O, Peacock EE, Sosik HM. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. In: Third Workshop on Fine-Grained Visual Categorization at CVPR 2015, 2015.

62. Munir S, Tran L, Francis J, Shelton C, Singh Arora R, Helsing C, Quintana M, Krishnan Prakash A, Rowe A, Berges M. Fine grained Occupancy estimatoR using Kinect on ARM Embedded Platforms. In: BuildSys 17 Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments]; 2017.
63. Xu L, Veeramachaneni K. Synthesizing Tabular Data using Generative Adversarial Networks. ArXiv, vol. abs/1811.11264; 2018.
64. Quintana M, Wai Tham K, Schiavon S, Miller C. Balancing thermal comfort datasets: We GAN, but should we? In: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation; 2020.
65. dos Santos Tanaka FHK, Arahna C. Data Augmentation Using GANs. In: Proceedings of Machine Learning Research XXX; 2019. p. 1–16.
66. Smith JW, Everhart J, Dickson W, Knowler W, Johannes R. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In: Proc Annu Symp Comput Appl Med Care, pp. 261–265, 1988.
67. Dheeru D, Graff C. UCI machine learning repository. Irvine: University of California, Irvine, School of Information and Computer Sciences, 2017.
68. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontepi G. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems; 2017. p. 1–14.
69. Deepshikha K, Naman A. Removing Class Imbalance using Polarity-GAN: An Uncertainty Sampling Approach. Conference on Computer Vision and Pattern Recognition; 2020.
70. Lopez Chau A, Li X, Yu W, Cervantes J, Mejia-Alvarez P. Border samples detection for data mining applications using non convex hulls. Mexican International Conference on Artificial Intelligence; 2011. p. 261–72.
71. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019.
72. Bishop MC. Pattern recognition and machine learning. New York: Springer Science+Business Media, LLC; 2006.
73. Kingma DP, Welling M. Auto-encoding variational bayes. In: International Conference on Learning Representations; 2013.
74. Srivastava A, Valkov L, Russell C, Gutmann MU, Sutton C. Veegan: Reducing mode collapse in gans using implicit variational learning. In: Advances in Neural Information Processing Systems; 2017.
75. Jordon J, Yoon J, van der Schaar M. Pate-gan: Generating synthetic data with differential privacy guarantees. In: International Conference on Learning Representations; 2019.
76. Ren J, Liu Y, Liu J. EWGAN: Entropy-Based Wasserstein GAN for Imbalanced Learning. In: The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19); 2019.
77. Montahaei E, Ghorbani M, Baghshah MS, Rabiee HR. Adversarial Classifier for Imbalanced Problems. *arXiv,* vol. abs/1811.08812; 2018.
78. Schlegl T, Seebock P, Waldstein SM, Schmidt-Erfurth U, Langs G. Unsupervised Anomaly Detection withGenerative Adversarial Networks to GuideMarker Discovery. In: Information Processing in Medical Imaging; 2021.
79. Mizra B, Haroon D, Khan B, Padhani A, Syed TQ. Deep generative models to counter class imbalance: a model-metric mapping with proportionality calibration methodology. In: IEEE Access; 2015. p. 55879–97.
80. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In: IEEE International Conference on Computer Vision (ICCV), 2017; 2017.
81. Redford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In: International Conference on Learning Representations 2016; 2015.
82. He H, Bai Y, Garcia EA, Li S.ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: IEEE World Congress on Computational Intelligence; 2008.
83. Osindero S, Mirza M. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs, stat]; 2014.
84. Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl. 2017;91:464–71.
85. Salimans T, Goodfellow I, Zaremba W, Radford A, Chen X. Improved Techniques for training GANs. In: Advances in Neural Information Processing Systems (NIPS); 2016.
86. Vacarri I, Orani V, Paglialonga A, Cambiaso E, Mongelli M. A Generative Adversarial Network (GAN) Technique for nternet of Medical Things Data. Sensors. 2021;21:3726.
87. Park N, Mohammadi M, Gorde K, Jajodia S, Park J, Kim Y. Data Synthesis based on Generative Adversarial Networks.. In: 44th International Conference on Very Large Data Bases 2018; 2018.
88. Okerinde A, Shamir L, Hsu W, Theis T, Nafi N. eGAN: Unsupervised approach to class imbalance using transfer learning. In: 2021 The 19th International Conference on Computer Analysis of Images and Patterns (CAIP); 2021.
89. Khoshgoftaar TM, Seiffert C, Van Hulse J, Napolitano A, Folleco A.Learning with limited minority class data. In: Sixth International Conference on Machine Learning and Applications (ICMLA 2007); 2007.
90. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Adv Neural Inf Proces Syst. 2017;30:8.
91. Sajjadi MS, Bachem O, Lucic M, Bousquet O, Gelly S.Assessing Generative Models via Precision and Recall. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal; 2018.
92. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

Publisher's Note