International Journal of Population Data Science

Journal Website: www.ijpds.org

Federated learning for generating synthetic data: a scoping review

Claire Little[1,*], Mark Elliot[2], and Richard Allmendinger[3]

Submission History

Submitted:	19/04/2023
Accepted:	18/09/2023
Published:	31/10/2023

[1]Cathie Marsh Institute for Social Research, School of Social Sciences, University of Manchester, Oxford Road, M13 9PL, Manchester, UK
[2]Department of Social Statistics, School of Social Sciences, University of Manchester, Oxford Road, M13 9PL, Manchester, UK
[3]Alliance Manchester Business School, University of Manchester, Oxford Road, M13 9PL, Manchester, UK

Abstract

Introduction
Federated Learning (FL) is a decentralised approach to training statistical models, where training is performed across multiple clients, producing one global model. Since the training data remains with each local client and is not shared or exchanged with other clients the use of FL may reduce privacy and security risks (compared to methods where multiple data sources are pooled) and can also address data access and heterogeneity problems. Synthetic data is artificially generated data that has the same structure and statistical properties as the original but that does not contain any of the original data records, therefore minimising disclosure risk. Using FL to produce synthetic data (which we refer to as "federated synthesis") has the potential to combine data from multiple clients without compromising privacy, allowing access to data that may otherwise be inaccessible in its raw format.

Objectives
The objective was to review current research and practices for using FL to generate synthetic data and determine the extent to which research has been undertaken, the methods and evaluation practices used, and any research gaps.

Methods
A scoping review was conducted to systematically map and describe the published literature on the use of FL to generate synthetic data. Relevant studies were identified through online databases and the findings are described, grouped, and summarised. Information extracted included article characteristics, documenting the type of data that is synthesised, the model architecture and the methods (if any) used to evaluate utility and privacy risk.

Results
A total of 69 articles were included in the scoping review; all were published between 2018 and 2023 with two thirds (46) in 2022. 30% (21) were focussed on synthetic data generation as the main model output (with 6 of these generating tabular data), whereas 59% (41) focussed on data augmentation. Of the 21 performing federated synthesis, all used deep learning methods (predominantly Generative Adversarial Networks) to generate the synthetic data.

Conclusions
Federated synthesis is in its early days but shows promise as a method that can construct a global synthetic dataset without sharing any of the local client data. As a field in its infancy there are areas to explore in terms of the privacy risk associated with the various methods proposed, and more generally in how we measure those risks.

Keywords
synthetic data; federated learning; review; data utility; data confidentiality

*Corresponding Author:
Email Address: claire.little@manchester.ac.uk (Claire Little)

https://doi.org/10.23889/ijpds.v8i1.2158

Introduction

Rationale

The ability to share data is important for transparency, research, and policy development. To ensure confidentiality, Statistical Disclosure Control (SDC) [1] can be applied to alter or remove disclosive information from data, making it safer for release. However, SDC methods can distort relationships in data, reducing its usefulness [2, 3]. Even with SDC applied there is still residual risk and if naively carried out it is possible to identify individuals in supposedly anonymous data [4, 5]. An alternative privacy-preserving method is data synthesis [6, 7] which uses models based on an original dataset to generate artificial data with the same structure and statistical properties, but which does not contain any of the original records. Synthetic data can be used where access to the original data is restricted or the data are unavailable for release due to privacy constraints. It may also be used for data augmentation, that is where extra data is created to add to an already existing dataset (for instance, large training datasets can be required when training a machine learning (ML) model). Where data is safeguarded, the application and approval processes to acquire access can be lengthy, taking months or even years and potentially delaying research analysis [8]. In these situations, synthetic versions of the data can allow researchers to plan analysis or test code whilst awaiting access to the real data.

Whilst synthetic data should present extremely low disclosure risk, as the risk of reidentification is not meaningful (since it does not contain any "real" data), there is still a risk of attribution disclosure [9]. This can occur where some attribute can be associated with a particular equivalence class; for example, the synthetic data might reveal that all females aged over 80 in a particular geographical area have dementia. There are therefore usually two competing objectives when producing synthetic data: high data utility (i.e., ensuring that the synthetic data is useful, with a distribution close to the original) and low disclosure risk. Balancing this trade-off can be difficult as, in general, reducing disclosure risk comes at a cost in utility.

Methods to generate synthetic data include statistical methods [10, 11] and deep learning (DL) methods based on neural networks (NNs) such as Generative Adversarial Networks (GANs) [12], Variational Autoencoders (VAEs) [13], large language models [14] and diffusion models [15, 16]. DL methods have been used extensively for image synthesis (and tend to deal with homogeneous numerical data) but there is growing interest in their use for the generation of tabular[1] data with methods being adapted for this purpose (i.e., to deal with predominantly categorical data, for example [17, 18]). Kokosi et al. [19] note that the most effective way of synthesising tabular data is not yet known, particularly where the datasets contain thousands or even millions of records and many variables, as is typical of administrative data.

GANs are a generative method that use NNs to model the distribution of some data. Broadly, a GAN aims to probabilistically describe how a dataset is generated, allowing new data to be generated by sampling from the model. Figure 1 shows the structure of a typical GAN, where two NN models are trained: a generative model that captures the data distribution and generates new data samples, and a discriminative model that aims to determine whether a sample is from the model distribution or the (real) data distribution. The models are trained together in an adversarial zero-sum game, such that the generator's goal is to produce data samples that fool the discriminator into believing that they are real, and the discriminator's goal is to determine which samples are real and which are fake. Training is iterative with (ideally) both models improving over time (in terms of capability) but with the overall system goal being a situation where the discriminator can no longer distinguish which data is real or fake (in terms of outcome).[2]

Federated Learning (FL) [20], is a decentralised approach to training statistical models, where the model hosted on a server is collaboratively trained using data from multiple clients, without the clients transmitting or exchanging the raw data.[3] Kairouz et al. [21] describe a typical FL training process using the FedAvg (Federated Averaging) algorithm [20], as involving a server orchestrating the training process by repeating the following steps:

1. **Client selection**: server samples from a set of clients who meet eligibility requirements (e.g., mobile phone is plugged in and connected to wi-fi).
2. **Broadcast**: the selected clients download current model weights and training program.
3. **Client computation**: each selected client locally runs and updates the model (training on their own local data).
4. **Aggregation**: server collects client updates and aggregates them.
5. **Model update**: server updates the global shared model based on the aggregated update from the clients.

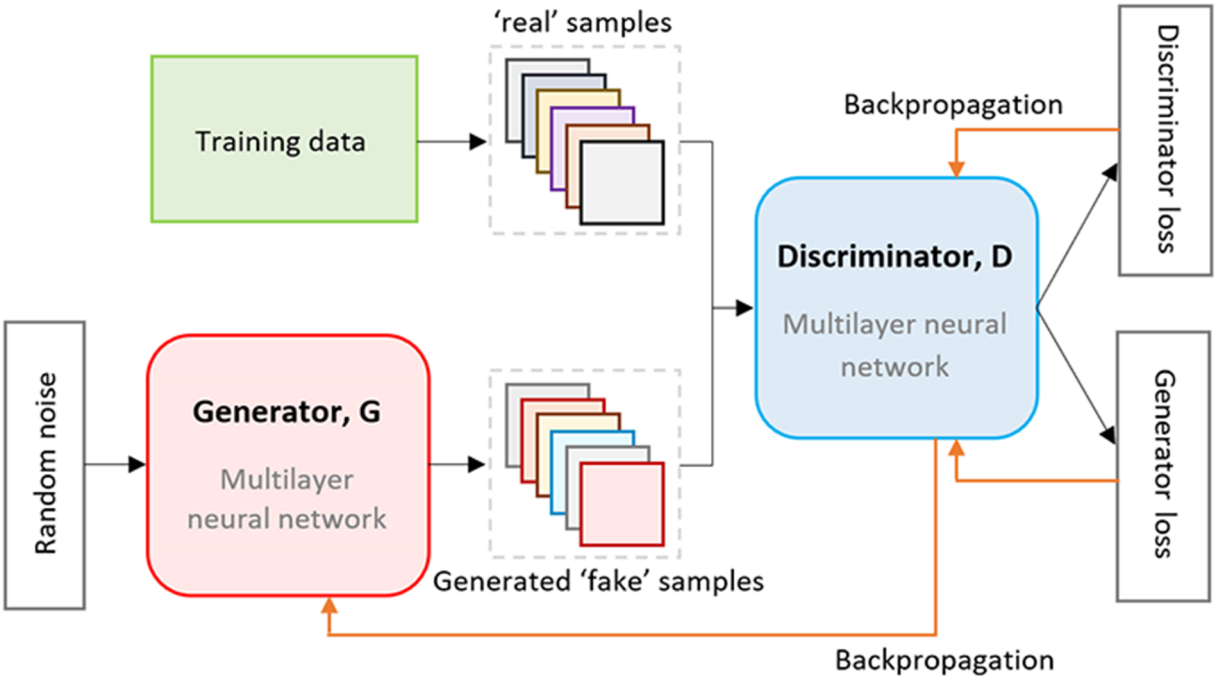
Training would usually continue until some stopping condition is met, for example, when a target number of training rounds have been performed. FL is commonly used to train NN models, with the weights and parameter settings sent to the central server. Since the central server does not access the local data during the training process, privacy and security risks are reduced (over alternative methods where the data would need to be transmitted and stored centrally). Challenges in FL include dealing with non-iid (non-independent and identically distributed) data on the clients and ensuring that the privacy of client data is truly preserved. Data can be non-iid in diverse ways, there could be different distributions or features in the client datasets, or clients may have differing amounts of data, all of which can affect the performance of the overall model

[1] Following [88], we refer to tabular data in the context of ML, as structured data comprising rows (e.g., individuals, cases) and columns (e.g., features, variables) that may contain mixed feature types (such as categorical, numerical, and ordinal).

[2] From a data synthesis perspective GANs are interesting in that the generative model does not access the original (or training) data at all and starts off with only noise as input. In theory this might reduce disclosure risk.

[3] Following the practice used in the papers we have reviewed, throughout this paper, we refer to each individual participating device as a client (the devices could be as diverse as a data server or a mobile phone, for example).

Figure 1: Structure of a typical GAN (Generative Adversarial Network)



NNs contain multiple layers of weighted nodes and backpropagation (backward propagation of errors) involves feeding the errors (or gradient of the loss function) back through the layers in order to adjust the weights for the next round of training.

[22]. As for privacy protection, it is possible that information about the raw data could still be leaked even where data is not sent, for instance, an adversary may be able to make inferences about a client's training data by knowing the previous model and the current model update supplied by that client [21]. Methods such as encryption or Differential Privacy (DP) [23] can be used to counter this risk, but it is important to be aware of the types of attack that could occur and where information could potentially be leaked.

The initial emphasis of FL was on mobile and edge device applications [20, 24, 25] where there could be many massively distributed clients with potentially limited communication, unbalanced data and/or different computational and storage capabilities [26], but interest has increased in its use in other applications [21]. For example, FL can allow cross-organisational collaboration for the training of models; potentially allowing for a (privacy-preserving) use of data that would never normally be shared or combined outside of an organisation, or institution, etc. In the context of healthcare, FL has the potential to improve medical care, allowing the creation of accurate, robust models without the need to exchange or centralise sensitive medical data [27, 28].

Using FL to produce synthetic data allows the combination of data from multiple clients, in a form that reflects real data, whilst minimising disclosure risk. It could allow the combination of information in datasets where those datasets would be unlikely to be linked in the traditional sense, thereby producing opportunities to access unique data that is potentially more diverse, and richer, than each client's synthetic dataset alone. That is, using FL to produce synthetic data may produce a dataset that is more representative of the overall (or combined) distribution than any clients own individual dataset could be, without the need for data sharing. As an example, if multiple hospitals would like to share their private patient data (pertaining to a medical trial, for instance) in order to have a larger sample, but they are restricted from doing this due to privacy constraints, an alternative then would be to use FL to produce synthetic data.

Objectives

FL is a relatively new research field, and it is unclear which methods are best suited for federated synthesis.[4] This paper aims to identify work to date, what type of data is generated, which model architecture (for example, GANs) is used to generate the data, which (if any) produce usable high utility data with low privacy risk, and how the privacy risk is measured (for example, is it a part of the algorithm, or ad-hoc). Since population data is typically likely to be in a tabular format, we perform a more detailed analysis on those methods producing tabular data. To meet these objectives, a scoping review was performed to systematically map the research done in this area and to identify any existing gaps in knowledge.

The review aims to investigate federated synthesis by: 1) identifying and describing the literature; 2) discussing open questions and trends; and 3) considering the benefits and challenges of federated synthesis. This scoping review will serve to inform the next stages of research and to our knowledge, is the first review undertaken on the use of FL for data synthesis.

[4]We define "federated synthesis" as the process of generating synthetic data using FL.

Methods

A scoping review was conducted to systematically map and describe the published literature on the use of FL to generate synthetic data, following the reporting guidelines of the PRISMA extension for scoping reviews (PRISMA-ScR) [29] and the five-stage framework defined by Arksey and O'Malley [30]. This included defining the research questions, identifying relevant studies, selecting studies, charting the data, and summarising and reporting the findings.

The search strategies and data charting form were drafted by Little and refined through discussion with the research team. Little performed the search, selection of sources of evidence and appraisal and synthesis of results, all with feedback from the research team.

Eligibility criteria

Articles from peer-reviewed journals were included together with conference and arXiv preprint papers. arXiv papers were included because this is a fast-moving field, and it is common for research to first appear on the platform. We recognise that because arXiv is not peer-reviewed, the quality of papers on the platform may be lower but since this review aims to include all research that is being undertaken it was deemed important to err on the side of inclusion. To be included, all research must be written in English and published between 2010 and January 2023. Since the term "federated learning" was first introduced in 2016 [21] it is unlikely there would be results before that date, but a longer timeline was used for the sake of thoroughness.

The focus of much of the FL literature is not data synthesis, but can involve goals that generally include producing more efficient algorithms to improve overall FL performance (where performance might be measured as the accuracy of a global predictive model, for example), or solving problems of non-iid data on the client side. Synthetic data is often produced as part of these models however and may be used to improve the training of an ML classifier by providing extra data, though it is not necessarily the final output of the model. The data produced by some of these models is referred to as augmented data. Whilst augmented data can simply be the original training data with transformations applied to it, it can also be produced using data synthesis methods, and therefore of interest in this review. Since, given the newness of the field we aim to be inclusive, a fairly broad set of search terms was used and variations of the terms "synthetic" and "augmented" were searched for alongside "federated learning". Papers were excluded if they did not use FL and/or did not generate new (synthetic or augmented) data.

Information sources and search

To identify potentially relevant documents, on January 31st 2023, the following bibliographic databases were searched: ACM, ScienceDirect, Web of Science, IEEE Xplore, Scopus and arXiv. The search strategy involved searching the title, abstract and keywords for the search terms (and variations of) "federated learning", "synthetic" and "augmentation", as listed in Table 1. The Google Scholar search engine was also used, but since it is not possible to focus the search on a particular field (e.g., title), the results were necessarily broader – therefore whilst all results were screened, only the top 50 were reviewed in depth since the relevance declined (and there were no results from Google Scholar that were not also identified in the bibliographic databases). The inclusion of papers identified whilst reviewing the primary sources and background literature was permitted if they satisfied the inclusion criteria; the cut-off date for this was February 13th 2023, after which no further articles were added. The results from each source were saved in a csv file.

Selection of sources of evidence

Once all sources were identified via search, the process involved firstly removing duplicates, then reading the abstract of each remaining article to apply the exclusion criteria. Articles were excluded if they did not use FL (e.g. [31] which discussed image synthesis but did not reference FL, although it is listed as a keyword in the paper), or if they used FL but did not involve the creation of new data (e.g. [32] which tested an FL approach on synthetic data but did not generate the synthetic data). Articles that were earlier versions of a newer work (that was already included) were also excluded (e.g. [33] was excluded as it was an earlier conference version of [34]).

Data charting, appraisal, and synthesis of results

After the exclusions were applied, all sources were loaded into the Mendeley reference manager and then the full article was reviewed. The data was captured using an Excel spreadsheet and included the following information:

- Author, title, publication title, year, DOI, URL, source, author affiliations, author country
- Goal of the study, type of data, application domain
- Method for generating data, FL algorithm, URL for code repository
- How study was evaluated, privacy metrics, privacy concerns
- Outcome of the study

The data charting variables were specified in advance; however, further variables were added during the capture process as became necessary. Extra variables were added to capture the distinct types of GAN architecture that were used, and to allow the inclusion of extra detail around the type of data that the clients and server exchange.

The papers were then grouped and summarised by: the overall aim (e.g., to generate synthetic data, or improve FL performance); the type of data they produced (e.g., image, tabular); whether they measured the associated privacy risk of generated data; and whether they send data (other than model parameters) to the server. The groupings were not independent across these dimensions, and for the data type a paper could contain a method that produces more than one type of data. Papers were also grouped by the model architecture (e.g., GAN) that they employed to generate the data and variations within this were summarised.

Table 1: The databases and search terms used

Database	Search terms
ACM Digital Library	[[Title: "federated learning"] AND [Title: synth*]] OR [[Title: "federated learning"] AND [Title: augmen*]] OR [[Abstract: "federated learning"] AND [Abstract: synth*]] OR [[Abstract: "federated learning"] AND [Abstract: augmen*]] OR [[Keywords: "federated learning"] AND [Keywords: synth*]] OR [[Keywords: "federated learning"] AND [Keywords: augmen*]] AND [E-Publication Date: (01/01/2010 TO 31/01/2023)]
arXiv	Query: date_range: from 2010-01-01 to 2023-01-31; include_cross_list: True; terms: AND title="federated learning" AND synth*; OR title="federated learning" AND augmen*; OR abstract="federated learning" AND synth*; OR abstract="federated learning" AND augmen*
IEEEXplore	("Document Title":"federated learning") AND ("Document Title":augmen*) OR ("Abstract":"federated learning") AND ("Abstract":augmen*) OR ("Author Keywords":"federated learning") AND ("Author Keywords":augmen*) Year: 2010–2023 ("Document Title":"federated learning") AND ("Document Title":synth*) OR ("Abstract":"federated learning") AND ("Abstract":synth*) OR ("Author Keywords":"federated learning") AND ("Author Keywords":synth*) Year: 2010–2023
ScienceDirect	Title, abstract or author-specified keywords: ("federated learning" AND augmentation) OR ("federated learning" AND augment) OR ("federated learning" AND synthetic) OR ("federated learning" AND synthesis) Year: 2010–2023
Scopus	TITLE-ABS-KEY ("federated learning" AND (synth* OR augmen*)) AND PUBYEAR > *2009*
Web of Science	(TI=(("federated learning" AND augmen*) OR ("federated learning" AND synth*))) OR (AB=(("federated learning" AND augmen*) OR ("federated learning" AND synth*))) OR (AK=(("federated learning" AND augmen*) OR ("federated learning" AND synth*))) Year: 2010-01-01 to 2023-01-31
Google Scholar	"federated learning" AND (synth* OR augmen*) Year: 2010-2023

Results

The PRISMA flow diagram (Figure 2) details the literature database and website search and article selection process. When all duplicates were removed and exclusions were applied, 69 articles remained.

Characteristics of the included papers (n = 69)

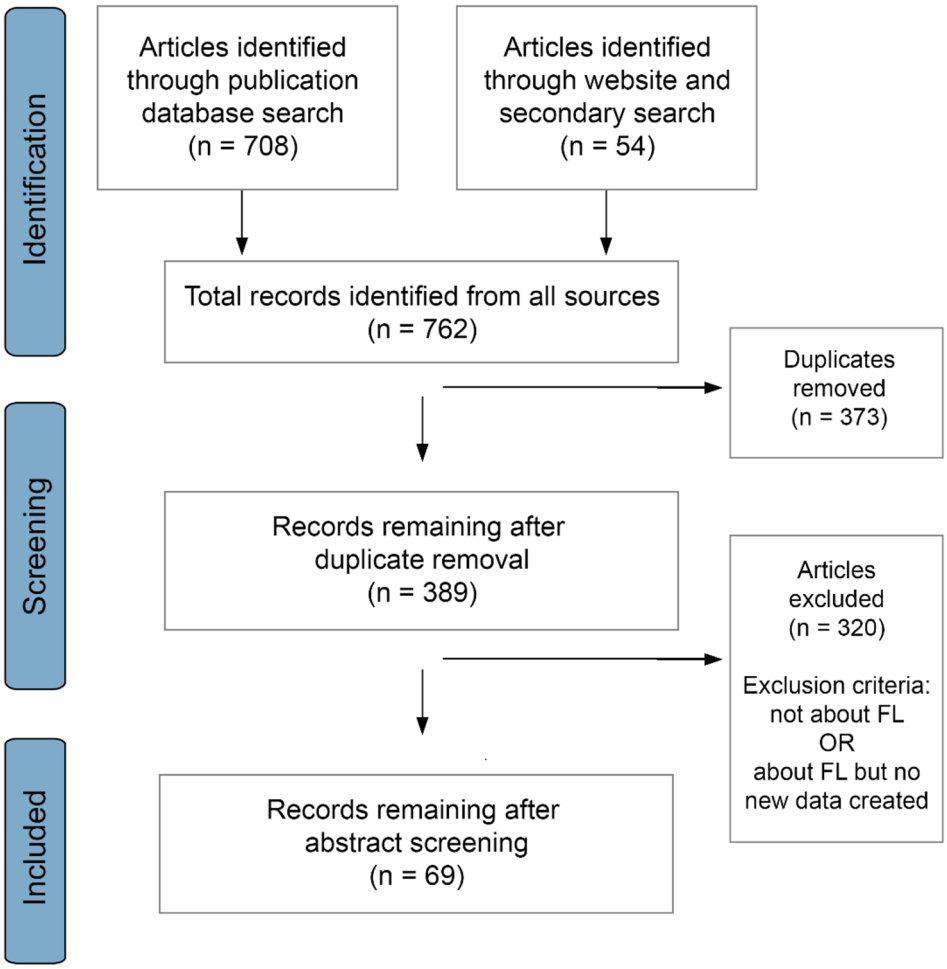
Figure 3 details the key characteristics of the selected papers. Two thirds (46, 67%) were published in 2022 suggesting a growing interest in FL. Just under half (32, 46%) of the papers did not specify a particular domain, whereas just over a quarter (19, 28%) were in the medical domain focussing predominantly (16 out of 19) on image data. Nine (13%) papers focussed on the Internet of Things (IOT), mobile or cloud services papers and six (9%) focussed on intrusion or anomaly detection. A third (23, 33%) of all papers originated in China (including Hong Kong), that is, the first author was based there. The USA (12, 17%) accounted for the next biggest group, followed by Germany (6, 9%). There was a mix of journal (23, 33%), conference (22, 32%) and arXiv (24, 35%) papers, and whilst nearly three quarters (51, 74%) of the papers were authored by academics, just over a fifth (15, 22%) were a collaboration between industry and academia. Supplementary Appendix 1 contains the data charting form with the variables and information extracted from the papers.

Categorising the papers

The 69 papers were grouped into four categories by the purpose for generating new data and the overall goal of the research:

- *Federated synthesis*: 21 (30%) papers had a goal of generating synthetic data, that is, the output was data that had been generated by the model. This is the group of central interest since these were methods designed specifically for generating synthetic data.
- *Improving FL using augmentation*: 29 (42%) papers generated data for the purpose of improving the FL process, that is, the data was used within the model (rather than as an output) to augment existing data. Whilst "Improving FL" is a broad categorisation, the papers in this group use data augmentation to mitigate the problem of non-iid data on the client side and improve the overall FL performance.
- *Developing FL*: 11 (16%) papers were categorised as developing FL further, this included a federated clustering framework [35] and seven papers [36–42] where clients share synthetic data with the server, rather than model parameters/weights (this might allow quicker model training and reduce communication costs).
- *Adversarial attacks and detection*: 8 (12%) papers focussed on attacks on FL, or the use of FL for intrusion

Figure 2: PRISMA flow diagram for the literature database and website search



or anomaly detection. They generated data to augment the models, although Huang et al. [43] demonstrated generating malicious data to poison the model.

These groups will be referred to in the following analysis.

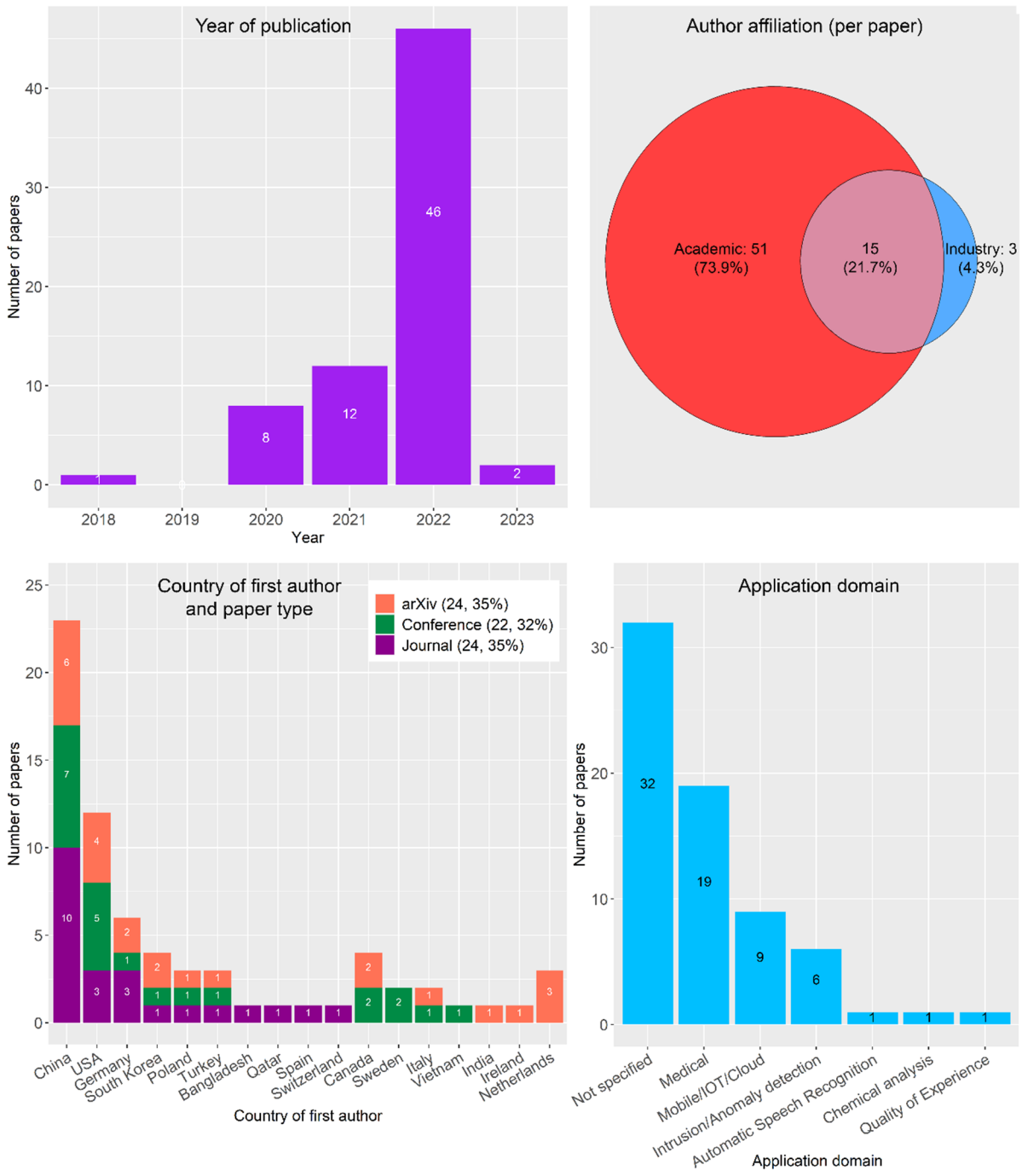
What type of data is generated?

The type of data was determined either by the papers explicitly stating it, or by the types of data that the method was evaluated on. Note that where a method could generate multiple types of data (e.g., image and text), this was not simultaneous, but as separate runs of the model. Overall, 72% (50) of the papers dealt with image data, 26% (18) used tabular data, 6% (4) used text data, 1% (1) used audio data and 1% (1) time series. There were overlaps in that six papers [35, 36, 40, 44–46] mentioned the ability to deal with multiple data types, such as image and text (not at the same time), and one paper [47] did not describe the data used. Figure 4 plots the different data types by the goal of research as described in the previous section. The papers in the *Federated synthesis* group deal with either image or tabular data ([45] could generate both), whereas papers in the *Improving FL using augmentation* group predominantly generate image data (with smaller groups generating tabular, text or audio data). The *Developing FL* group focus mostly on image data whereas the *Adversarial attacks and detection* group focus mostly on tabular data.

The methods used to generate image data can be different than those used for other data types. Where image data was augmented, this could simply involve transformations. That is, synthetic data generation methods are not used and transformations (such as cropping, rotation, blurring, etc.) are performed on the images – this was the case for seven papers [48–54], whilst [55–57] used transformations and GAN or VAE, and one paper [58] suggested augmenting data by using Google's image search function.

There can be problems associated with dealing with tabular data when using deep learning methods such as GANs, as they were designed initially to deal with numerical data and must be adapted for categorical data. Zhao et al [18] state that in general, when attempting to synthesise tabular data using a GAN it is important to know the global data properties (such as all possible distinct values that a categorical variable can take) so that the inputs and outputs of the models will be setup correctly. However, to do this would require sharing knowledge about the data by sending information to the central server about each client's data distribution, which could present a disclosure risk (for instance, if this was medical data, simply knowing that a certain number of patients at a particular site have a disease may be disclosive). Four of the papers [59–62]

Figure 3: Characteristics of included papers (n = 69)



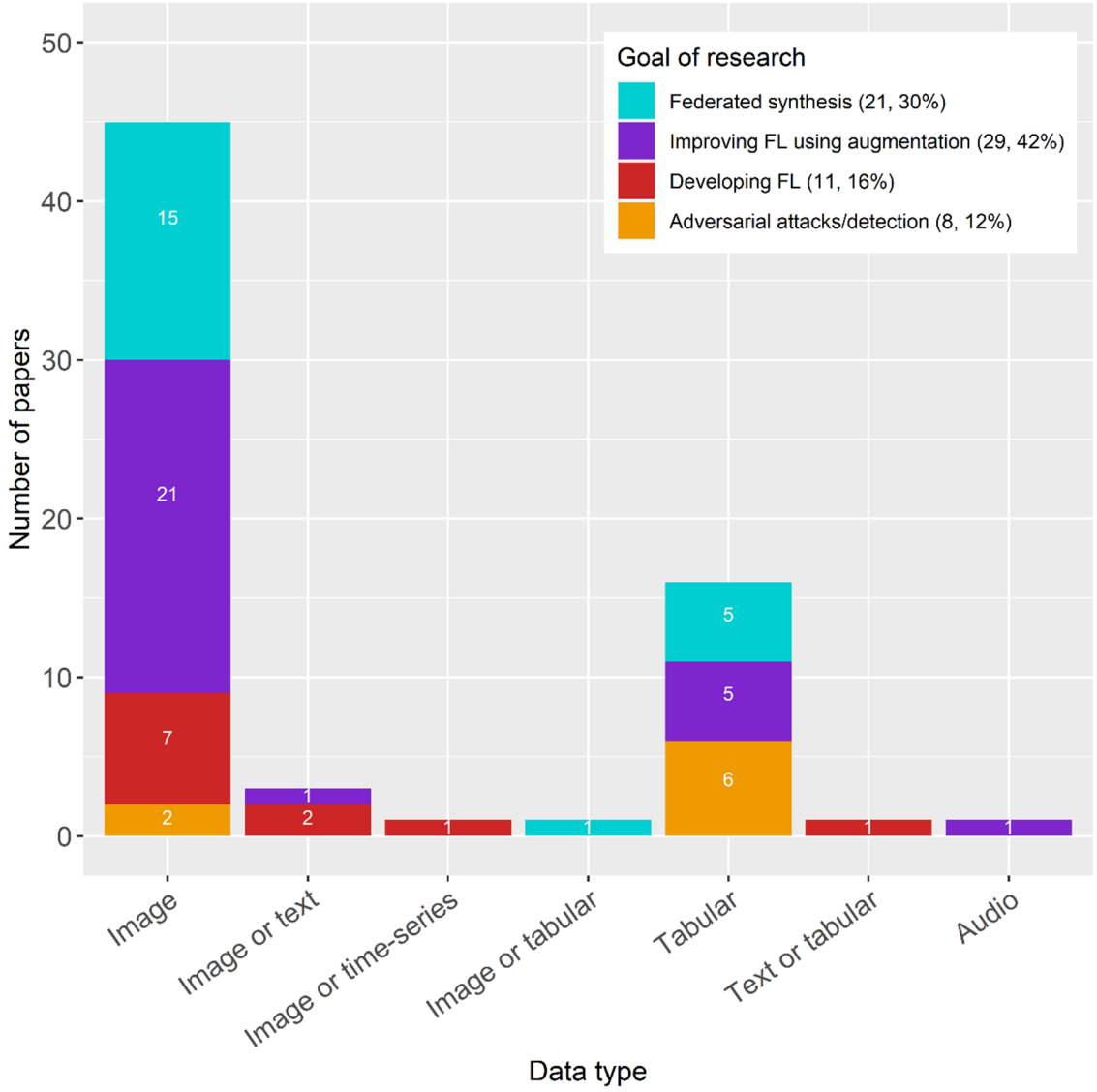
dealing with tabular data did send information about the client data properties to the server, however the other twelve papers did not share such information, which suggests it may be possible to produce tabular data using these methods without risking information disclosure.

What methods have been used for federated synthesis?

Since the primary focus of this review is to determine what methods are used for federated synthesis, Table 2 details the 21 papers that generate synthetic data using FL, i.e., those in the *Federated synthesis* group. All but two use GANs to generate the synthetic data and just under half (10) of the papers are in the medical domain. Nine of the papers implement Differential Privacy (DP) with the majority achieving this by clipping the gradients and adding noise to the GAN generator/discriminator. The GAN configuration column in Table 2 refers to Figure 5, which is explained in more detail in the following section and identifies the different configurations of the GANs used on client and server devices (where applicable).

Of the 21 papers, 16 generated image data and 6 generated synthetic tabular data (Lomurno et al. [45] could generate image or tabular data). It is clear there is not currently a large body of work on federated synthesis, particularly for tabular

Figure 4: The type of data used by each paper, by the goal of the research



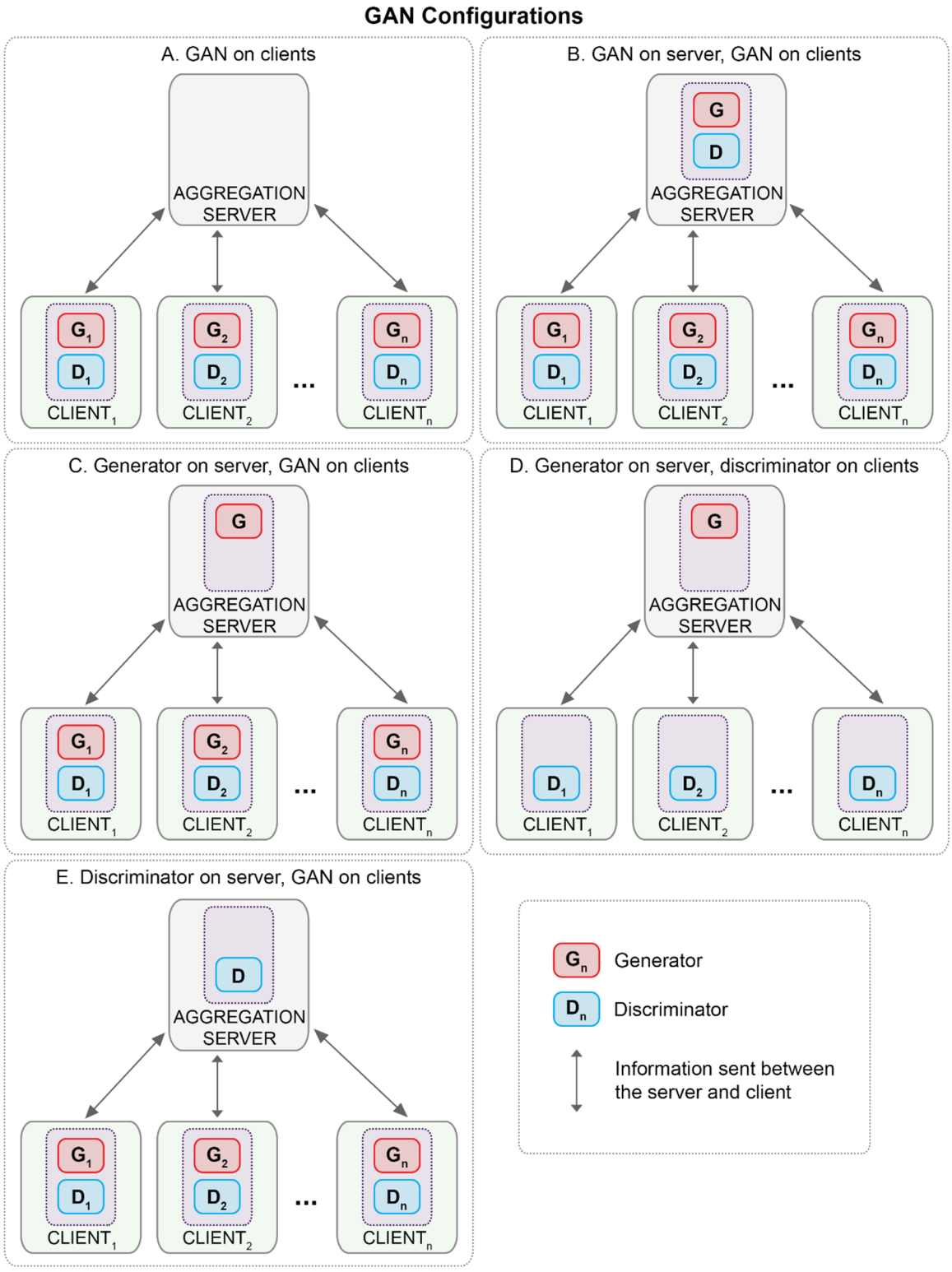
data. Since population data is more likely to be in a tabular format, the focus of this section is the research generating tabular data:

- *HT-Fed-GAN* [60] uses a federated variational Bayesian Gaussian mixture model and one-hot encoding to deal with problems caused by tabular data (such as multimodal distributions and imbalanced attributes), conditional GANs and DP. Each client had a local GAN, and the server aggregated the parameters. It was compared to another method (one not designed for tabular data) and outperformed this in terms of synthetic data utility (comparing predictive accuracy of classification/regression tests and cumulative distributions of attributes). *HT-Fed-GAN* was one of four papers included in this review to perform post-hoc privacy risk analysis on the generated data (using membership inference attack [77]) and its privacy risk was similar to the comparator model (whilst having better utility).

- *FDP-CTGAN* [66] - it should be noted this paper concentrates on presenting a version of CTGAN [17] (which is a GAN specifically designed to generate tabular data) with the addition of DP (DP-CTGAN), and a federated version is only briefly described. FDP-CTGAN synthesises tabular medical data, and its performance in terms of utility (predictive performance of a model trained on synthetic data but tested on real data) is compared to other (non-FL) methods; the federated version is outperformed (in terms of, the Area Under Receiver Operating Curve and Area Under the Precision-Recall curve metrics) in all but one case by CTGAN and DP-CTGAN.

- SGDE [45] presents a different way to generate synthetic data by having each client train a data generator locally (VAE are used) using DP; each client sends their generator to the server, but they are not aggregated or combined (as is typical in FL). In the final phase each client can access the set of generators stored on the

Figure 5: The different GAN (Generative Adversarial Network) configurations used on the server and client devices, for federated synthesis



server and use some/all of these to generate synthetic data locally. This method is not compared to others, but to a local and federated version, with the synthetic data version outperforming those overall, in terms of prediction accuracy, AUC and F1 score.

- Weldon et al. [74] generate synthetic electronic health records (EHR) using federated GANs and tested their method by constructing the scenario of multiple hospitals, each with unique data silos, training a GAN locally and combining their parameters into one central GAN, which generated the synthetic patient data. Whilst not compared to other methods, the synthetic data utility was evaluated (comparing predictive diagnosis accuracy to real data, using coefficient of determination (R^2) and RMSE (root mean squared error)) and domain

Table 2: Details of the 21 papers that perform federated synthesis (where DP is Differential Privacy, GAN is Generative Adversarial Network, and GAN configuration refers to Figure 5)

Lead Author, Year	Application domain	Data type	Model name	Synthetic data generation method	GAN configuration (if applicable)	Use DP
Behera, 2022 [63]	Not specified	Image	FedSyn	GAN	A	Yes
Dalmaz, 2022 [64]	Medical	Image	pFLSynth	GAN	C	No
Duan, 2022 [60]	Not specified	Tabular	HT-Fed-GAN	GAN	A	Yes
Elmas, 2022 [65]	Medical	Image	FedGIMP	GAN	C	No
Fang, 2022 [66]	Medical	Tabular	FDP-CTGAN	GAN	A	Yes
Han, 2020 [67]	Medical	Image	-	GAN	B	No
Li, 2020 [68]	Medical	Image	-	GAN	B	No
Lomurno, 2022 [45]	Not specified	Image, Tabular	SGDE	VAE	-	Yes
Mei, 2022 [69]	Not specified	Image	VPFL	GAN	A	No
Pfitzner, 2022 [70]	Not specified	Image	DPD-fVAE	VAE	-	Yes
Qu, 2020 [57]	Medical	Image	TDGAN	GAN, Transformations	D	No
Rajotte, 2021 [71]	Medical	Image	FELICIA	GAN	E	No
Triastcyn, 2020 [72]	Not specified	Image	FedGP	GAN	C	Yes
Wang, H, 2021 [73]	Mobile/IOT/Cloud	Image	P-FedGAN	GAN, AE	A	Yes
Wang, J, 2022 [55]	Medical	Image	FedMed-ATL	GAN, Transformations	C	No
Weldon, 2021 [74]	Medical	Tabular	-	GAN	B	No
Welten, 2022 [75]	Not specified	Image	synRainGAN	GAN	A	No
Xie, 2022 [76]	Medical	Image	FedMed-GAN	GAN	C	Yes
Xin, 2022 [34]	Not specified	Image	FL-GAN	GAN	B	Yes
Zhao, 2021 [61]	Not specified	Tabular	Fed-TGAN	GAN	A	No
Zhao, 2023 [62]	Not specified	Tabular	GTV	GAN	B	No

experts were used to rate the data on plausibility, finding no substantial difference between the synthetic and real data, although this was not statistically tested. The authors also performed a basic post-hoc privacy risk analysis, checking the synthesised data for records duplicated from the original data (there were none) or records that were highly similar (using Cosine similarity).

- Fed-TGAN [61] uses a GAN on each client and initialises the model by sending the statistical properties (categorical frequency distribution, and distribution of continuous variables) of each client's data to the server. This is performed in order to understand the global data distribution, but there may be disclosure risk in sending this type of data. Fed-TGAN is compared against other FL methods in terms of data quality (using average Jensen-Shannon divergence and Wasserstein distance), and it generally outperformed them.
- *GTV* [62] uses a GAN-based method for Vertical FL (where clients have unique data relating to the same individuals, rather than having similar data relating to different individuals). Like [61] information about the data distribution is collected in a "conditional vector" and shared in order to understand the global distribution. This conditional vector stores indices (corresponding to rows in the real data) that must be selected for training to balance the classes in a particular categorical column. Since this information is available to server and clients, it presents the risk of adversaries potentially reconstructing the categorical columns, although the method shuffles the training data after each iteration to mitigate this risk. GTV is not compared to other FL methods, but the performance impact of the number of FL clients is investigated; utility is compared (prediction accuracy and average Jensen-Shannon divergence, Wasserstein distance and correlation difference) and GTV produces results similar to a centralised (non-FL) method.

Which methods are used to generate the synthetic data?

Across all the included papers, whatever the goal (augmentation or synthesis), 62% (43) used deep learning methods to generate data, with the majority of these (36) being GAN-based methods (and four using VAE, two using both GAN and AE-based methods). Two papers [78, 79] used tree-based methods, and three others [80–82] used SMOTE (Synthetic Minority Oversampling Technique). Thirteen papers used image transformations, image search or the exchange of image statistics. As the most used method, the GANs were deployed in various architectures, as pictured in Figure 5, and are summarised here for the 21 papers in the *Federated synthesis* group. It should be noted that it was difficult to extract the exact architecture from some of the papers as the language was not always precise, for instance, a paper may refer to having a GAN on the server but only use the generator, whereas others may specifically mention using just a generator.

- **A. GAN on clients**: 7 papers [60, 61, 63, 66, 69, 73, 75] had a GAN on each client but not on the server. Usually, each client trains its GAN, sends the parameters to the server, which aggregates them and sends them back to the client, and so on. In this scenario once the FL model has finished training, the local GAN (generator) will generate the synthetic data on each client.

- **B. GAN on server, GAN on clients:** 5 papers [34, 62, 67, 68, 74] had a GAN on the server and on each client. Usually, the server GAN will initiate the process by sending initial parameters to each client, which trains its own GAN and then sends parameter updates back to the central server, which aggregates them and then the server GAN may train before sending parameters back to the clients. Two of these papers use sequential training, Weldon et al. [74] train each client in turn (each sends the parameters back to the central server after they train), whereas Xin et al. [34] use serial training, which involves passing the model parameters from client to client until each has participated, where it is then returned to the server (each time the model is passed on, the new client synthesises a small amount of data and adds it to their training data, in order to balance the data).

- **C. Generator on server, GAN on clients**: 5 papers [55, 64, 65, 72, 76] had a GAN on each client and a generator on the server. Usually, each client trains and sends generator parameters to the server, which aggregates them and sends them back, and so on. Once training has concluded each client can generate synthetic data, or, in theory, the server can use its generator to do this.

- **D. Generator on server, discriminator on clients**: 1 paper [57] had a generator on the server and a discriminator on the clients. Used for image generation, the central generator creates images, and the client discriminators evaluate them based on their own data. Rather than gradient updates, data (images) are exchanged between client and server.

- **E. Discriminator on server, GAN on clients:** 1 paper [71] had a GAN on the clients and a discriminator on the server. Each client trains its local GAN (using image data) and sends the synthetic data to the server discriminator, which has the same architecture as the client discriminators except for the activation of the final layer. The final layer would normally predict whether the image was "real" or fake, however it cannot do this as the central server has no access to the "real" data. The authors do not make it clear how the server discriminator operates, but it does train and feedback an error to the model.

Data utility and privacy risk

Overall, the performance of many of the models is more difficult to determine – 52% (36) did not compare their method against others (although 27 of the 36 compare variations of their own method). Most of those that did provide some kind of comparison against other methods found their framework or algorithm was at least comparable to other methods (i.e., there were no negative results). Evaluation was usually performed by considering the accuracy of a prediction task, with 58% (40) of papers using accuracy-based metrics, and 22% (15) using accuracy alongside other metrics (e.g., statistical similarity). For the 21 *Federated synthesis* papers (i.e., those who produced synthetic data as the output of the model) metrics evaluating the image quality were used on 11 of the 16 image methods, with prediction accuracy used on 9 papers (5 used multiple metrics). For the six tabular data methods, a combination of prediction accuracy (4) and statistical similarity (4) were used. Han et al. [67] and Weldon et al. [74] used domain experts to analyse their data.

When FL was first proposed, one of the core assumptions was that the data would stay on the local client devices and only model parameter updates would be sent to the global server; this was the selling point in terms of minimising disclosure risk. However, 39% (27) of the papers included in this review send some form of data (beyond parameters) to the central server. Ten papers send information about the data distributions to the server, with few appearing to consider the risk associated with this. Four papers mention sending data samples in order to compensate for data imbalance or to allow the central GAN to have access to data, with only one [83] referring to the samples being encoded. Thirteen papers send synthetic data to the server or other clients. Whilst 10 of the 27 papers that send data do mention using encryption or some form of DP, and 19 papers include a section discussing privacy risks, it is notable that some of the papers do not appear to have considered the potential risk associated with sending such data.

Overall, 67% (46) of the 69 papers do not mention using any form of privacy enhancing technology when sending updates to the server. However, 28% (19) do mention using DP, and a further four papers mention adding some form of noise or encryption to the updates sent to the server. Of those that used DP, many also tended to experiment with different values of ϵ (epsilon, which is a settable parameter to control the amount of noise added and hence the level of privacy afforded). All but four of the methods do not perform a post-hoc privacy measure on the data they create, although this may only be necessary if data is an output of the model. Overall, just over a half (37, 54%) of the included papers provide some discussion of privacy risk.

Discussion

This scoping review identified 69 papers that generated data using FL, of those 21 conducted federated synthesis. The field of FL is in its infancy, having begun in 2016, but it is rapidly growing, with a marked increase in literature in 2022. However, there is still only a small body of work on federated synthesis, particularly for tabular data, with most of the work carried out on image data.

Deep Learning methods, such as GANs, were the most used method for generating synthetic data, this may be partly because their popularity has risen in line with that of FL over the last few years. Another reason may be that neural network models in general are suited to the FL task because the weights/parameters of the models can easily be shared and aggregated. Also, the different configurations of GANs explored in this review (e.g., allowing a generator on the server and GANs on clients) allow for flexibility and experimentation. Since, in general, NN methods are less suited to the use of tabular data and tend to require adaptation, this may explain the relative lack of research in the federated synthesis of tabular data.

It was notable that a large portion of the included papers deviated from the original idea of FL [20, 21] by choosing to send potentially disclosive information (data other than parameters) to the server, and further research may be required to quantify the risk associated with the different types of data that are sent. The relative lack of engagement with privacy/confidentiality issues is concerning given that this is part of the raison d'etre for FL. It is possible that in some of these cases there is a naive assumption that simply because FL is being used, there is inherently less risk.

In general, there is a need for consensus on the metrics to measure the utility and risk (and a lack of post-hoc risk analysis generally) of the synthetic data produced, but this is also true of non-federated data synthesis. Recent work by Taub et al. [84, 85] have developed a basket approach for assessing utility of synthetic data - considering many different analytical properties of the data simultaneously. It is unclear how the different federated synthesis methods reviewed in this paper fare in terms of the risk-utility trade-off; it is possible to map this on the risk-utility map (for example, see [86, 87]). In general, post-hoc risk and utility measures for federated synthesis will need to be thought about differently, for the simple reasons that we don't have the original (combined) data to make comparisons to. So this is something that each client may need to perform separately (for example, a global synthetic dataset might be acceptable to client A in terms of disclosure risk but considered disclosive for client B); this is an area where further research is needed.

Conclusion

Federated synthesis is in its early days and it is unclear at this stage whether it is simply an interesting research problem or whether federated synthetic data may prove to have some practical value. However, the research reviewed here shows that the methodology has promise and warrants further research. The potential that we may be able to construct a global synthetic dataset without sharing any of the local client data is a sufficient prize to motivate future research effort.

As a field in its infancy there are areas to explore in terms of the privacy risk associated with the various methods proposed and the information that they transmit. In general, work on achieving greater consensus on how we measure both the risk and utility of synthetic data is required to effectively compare and evaluate methods.

Statement on conflicts of interest

The authors declare that they have no conflicts to report.

Ethics statement

This research article did not require ethical approval as the research is based on a review of published/publicly reported literature.

References

1. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, et al. Statistical Disclosure Control. John Wiley & Sons, Ltd; 2012. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118348239.

2. Purdam K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environ Plan A. 2007;39(5):1101–18. https://doi.org/10.1068/a38335

3. Drechsler J, Reiter JP. Sampling with synthesis: A new approach for releasing public use census microdata. J Am Stat Assoc. 2010;105(492):1347–57. https://doi.org/10.1198/jasa.2010.ap09480

4. Dwork C, Smith A, Steinke T, Ullman J. Exposed! A survey of attacks on private data. Annu Rev Stat Its Appl. 2017;4(1):61–84. https://doi.org/10.1146/annurev-statistics-060116-054123

5. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10(3069). https://doi.org/10.1038/s41467-019-10933-3

6. Rubin DB. Statistical Disclosure Limitation. J Off Stat. 1993;9(2):461–8. Available from: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.0

7. Little RJA. Statistical Analysis of Masked Data. J. Off. Stat. 1993:9(2):407–26. Available from: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

8. Taylor JA, Crowe S, Espuny Pujol F, Franklin RC, Feltbower RG, Norman LJ, et al. The road to hell is paved with good intentions: The experience of applying for national data for linkage and suggestions for improvement. BMJ Open. 2021;11(e047575). https://doi.org/10.1136/bmjopen-2020-047575

9. Taub J, Elliot M, Pampaka M, Smith D. Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: Domingo-Ferrer J, Montes F, editors. Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science(), vol 11126. Springer, Cham; 2018. https://doi.org/10.1007/978-3-319-99771-1_9

10. Nowok B, Raab GM, Dibben C. Synthpop: Bespoke creation of synthetic data in R. J Stat Softw. 2016;74(11). https://doi.org/10.18637/jss.v074.i11

11. Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In: Proceedings of SSDBM '17. New York, NY, USA: ACM; 2017. https://dl.acm.org/doi/10.1145/3085504.3091117

12. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc.; 2014. Available from: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

13. Kingma DP, Welling M. Auto-encoding variational bayes. In: International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. 2014. Available from: https://arxiv.org/abs/1312.6114.

14. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI blog [Internet]. 2019;1(8). Available from: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

15. Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. PMLR; 2015. p. 2256–65. Available from: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

16. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems 33. Curran Associates, Inc.; 2020. p. 6840–51. Available from: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

17. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

18. Zhao Z, Kunar A, Van der Scheer H, Birke R, Chen LY. CTAB-GAN: Effective Table Data Synthesizing. In: Proceedings of The 13th Asian Conference on Machine Learning. PMLR 157; 2021. p. 97–112. Available from: https://proceedings.mlr.press/v157/zhao21a.html.

19. Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, et al. An overview of synthetic administrative data for research. IJPDS. 2022;7(1). https://doi.org/10.23889/ijpds.v7i1.1727

20. McMahan HB, Moore E, Ramage D, Hampson S, Aguera y Arcas B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Artificial Intelligence and Statistics, PMLR 54. Fort Lauderdale, Florida; 2017. p. 1273–82. Available from: http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf.

21. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. Found Trends Mach Learn. 2021;14(1–2):1–210. https://doi.org/10.1561/2200000083

22. Hsieh K, Phanishayee A, Mutlu O, Gibbons PB. The Non-IID data quagmire of decentralized machine learning. In: 37th International Conference on Machine Learning. PMLR 119; 2020. p. 4387–98. Available from: https://proceedings.mlr.press/v119/hsieh20a.html.

23. Dwork C, Roth A. The algorithmic foundations of differential privacy. In: Foundations and Trends in Theoretical Computer Science. 2014;9(3-4):211–487. https://doi.org/10.1561/0400000042

24. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al. Practical Secure Aggregation for Federated Learning on User-Held Data. In: NIPS Workshop on Private Multi-Party Machine Learning. 2016. Available from: http://arxiv.org/abs/1611.04482.

25. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated Learning: Strategies for Improving Communication Efficiency. In: NIPS Workshop on Private Multi-Party Machine Learning. 2016. Available from: http://arxiv.org/abs/1610.05492.

26. Li T, Sahu AK, Talwalkar A, Smith V. Federated Learning: Challenges, methods and future directions. IEEE Signal Process Mag. 2020;37(3):50–60. https://ieeexplore.ieee.org/document/9084352

27. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digit Med. 2020;3(119). https://doi.org/10.1038/s41746-020-00323-1

28. Kumar Y, Singla R. Federated Learning Systems for Healthcare: Perspective and Recent Progress. In: Rehman, MHu, Gaber, MM, editors. Federated Learning Systems Studies in Computational Intelligence, vol 965. Springer, Cham; 2021. p. 141–56. https://doi.org/10.1007/978-3-030-70604-3_6

29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467–73. https://doi.org/10.7326/M18-0850

30. Arksey H, O'Malley L. Scoping studies: Towards a methodological framework. Int J Soc Res Methodol Theory Pract. 2005;8(1):19–32. https://doi.org/10.1080/1364557032000119616

31. Chen JW, Yu CM, Kao CC, Pang TW, Lu CS. DPGEN: Differentially Private Generative Energy-Guided Network for Natural Image Synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. p. 8377–86. https://doi.org/10.1109/CVPR52688.2022.00820

32. Zhan Y, Li P, Wu L, Guo S. L4L: Experience-Driven Computational Resource Control in Federated Learning. IEEE Trans Comput. 2022;71(4):971–83. https://doi.org/10.1109/TC.2021.3068219

33. Xin B, Yang W, Geng Y, Chen S, Wang S, Huang L. Private FL-GAN: Differential privacy synthetic data generation based on federated learning. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2020. p. 2927–31. https://doi.org/10.1109/ICASSP40776.2020.9054559

34. Xin B, Geng Y, Hu T, Chen S, Yang W, Wang S, et al. Federated synthetic data generation with differential privacy. Neurocomputing. 2022;468:1–10. https://doi.org/10.1016/j.neucom.2021.10.027

35. Yan J, Liu J, Qi J, Zhang ZY. Federated clustering with GAN-based data synthesis. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: https://arxiv.org/abs/2210.16524v1.

36. Hu S, Goetz J, Malik K, Zhan H, Liu Z, Liu Y. FedSynth: Gradient Compression via Synthetic Data in Federated Learning. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2204.01273.

37. Goetz J, Tewari A. Federated Learning via Synthetic Data. arXiv [Internet]. 2020 [cited 2023 Feb 2]. Available from: https://arxiv.org/abs/2008.04489.

38. Pennisi M, Salanitri FP, Palazzo S, Pino C, Rundo F, Giordano D, Spampinato, C. GAN Latent Space Manipulation and Aggregation for Federated Learning in Medical Imaging. In: Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health. DeCaF FAIR 2022. Lecture Notes in Computer Science, vol 13573. Springer, Cham; 2020. https://doi.org/10.1007/978-3-031-18523-6_7

39. Xiong Y, Wang R, Cheng M, Yu F, Hsieh C-J. FedDM: Iterative Distribution Matching for Communication-Efficient Federated Learning. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: https://arxiv.org/abs/2207.09653.

40. Zhou Y, Ma X, Wu D, Li X. Communication-Efficient and Attack-Resistant Federated Edge Learning with Dataset Distillation. IEEE Trans Cloud Comput. 2022;1–12. https://doi.org/10.1109/TCC.2022.3215520

41. Yan Z, Wicaksana J, Wang Z, Yang X, Cheng K-T. Variation-Aware Federated Learning With Multi-Source Decentralized Medical Image Data. IEEE J Biomed Heal Informatics. 2021;25(7):2615-2628. https://doi.org/10.1109/JBHI.2020.3040015

42. Zhu S, Qi Q, Zhuang Z, Wang J, Sun H, Liao J. FedNKD: A Dependable Federated Learning Using Fine-tuned Random Noise and Knowledge Distillation. In 2022 International Conference on Multimedia Retrieval (ICMR '22). Association for Computing Machinery; 2022. p185–193. https://doi.org/10.1145/3512527.3531372

43. Huang J, Zhao Z, Chen LY, Roos S. Blind leads Blind: A Zero-Knowledge Attack on Federated Learning. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2202.05877.

44. Jiang X, Hu H, On T, Lai P, Mayyuri VD, Chen A, et al. FLSys: Toward an Open Ecosystem for Federated Learning Mobile Apps. IEEE Trans Mob Comput. 2022. https://doi.org/10.1109/TMC.2022.3223578

45. Lomurno E, Archetti A, Di Perna L, Cazzella L, Samele S, Matteucci M. SGDE: Secure Generative Data Exchange for Cross-Silo Federated Learning. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2109.12062.

46. Yoon T, Shin S, Hwang SJ, Yang E. FedMix: Approximation of Mixup under Mean Augmented Federated Learning. In: International Conference on Learning Representations, ICLR 2021. 2021. Available from: https://openreview.net/pdf?id=Ogga20D2HO-.

47. Yan M, Chen B, Feng G, Qin S. Federated Cooperation and Augmentation for Power Allocation in Decentralized Wireless Networks. IEEE Access. 2020;8:48088–100. https://doi.org/10.1109/ACCESS.2020.2979323

48. Cetinkaya AE, Akin M, Sagiroglu S. Improving Performance of Federated Learning based Medical Image Analysis in Non-IID Settings using Image Augmentation. 2021 International Conference on Information Security and Cryptology (ISCTURKEY). IEEE; 2021. p. 69–74. https://doi.org/10.1109/ISCTURKEY53027.2021.9654356

49. de Luca AB, Zhang G, Chen X, Yu Y. Mitigating Data Heterogeneity in Federated Learning with Data Augmentation. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2206.09979.

50. Duan M, Liu D, Chen X, Liu R, Tan Y, Liang L. Self-Balancing Federated Learning with Global Imbalanced Data in Mobile Systems. IEEE Trans Parallel Distrib Syst. 2021;32(1):59–71. https://doi.org/10.1109/TPDS.2020.3009406

51. Hossen MN, Panneerselvam V, Koundal D, Ahmed K, Bui FM, Ibrahim SM. Federated Machine Learning for Detection of Skin Diseases and Enhancement of Internet of Medical Things (IoMT) Security. IEEE J Biomed Heal Informatics. 2022;27(2):835–841. https://doi.org/10.1109/JBHI.2022.3149288

52. Jia J, Mahadeokar J, Zheng W, Shangguan Y, Kalinli O, Seide F. Federated Domain Adaptation for ASR with Full Self-Supervision. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2022. p. 536–40. Available from https://www.isca-speech.org/archive/pdfs/interspeech_2022/jia22_interspeech.pdf.

53. Linardos A, Kushibar K, Walsh S, Gkontra P, Lekadir K. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. Sci Rep. 2022;12(1):1–12. https://doi.org/10.1038/s41598-022-07186-4

54. Wang T, Cao Z, Wang S, Wang J, Qi L, Liu A, et al. Privacy-Enhanced Data Collection Based on Deep Learning for Internet of Vehicles. IEEE Trans Ind Informatics. 2020;16(10):6663–72. https://doi.org/10.1109/TII.2019.2962844

55. Wang J, Xie G, Huang Y, Zheng Y, Jin Y, Zheng F. FedMed-ATL: Misaligned Unpaired Cross-Modality Neuroimage Synthesis via Affine Transform Loss. In: Proceedings of the 30th ACM International Conference on Multimedia. Association for Computing Machinery; 2022. p. 1522–31. https://doi.org/10.1145/3503161.3547762

56. Kim Y, Lee W. Distributed Raman Spectrum Data Augmentation System Using Federated Learning with Deep Generative Models. Sensors. 2022;22(24):9900. https://doi.org/10.3390/s22249900

57. Qu H, Zhang Y, Chang Q, Yan Z, Chen C, Metaxas D. Learn Distributed GAN with Temporary Discriminators. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12372. Springer, Cham; 2020. p. 175–92. https://doi.org/10.1007/978-3-030-58583-9_11

58. Jeong E, Oh S, Kim H, Kim S Iyun, Park J. Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data. arXiv [Internet]. 2018 [cited 2023 Feb 2]. Available from: https://arxiv.org/abs/1811.11479.

59. Duan S, Liu C, Han P, He T, Xu Y, Deng Q. Fed-TDA: Federated Tabular Data Augmentation on Non-IID Data. arXiv [Internet]. 2022 [cited 2023 Feb 5]. Available from: http://arxiv.org/abs/2211.13116.

60. Duan S, Liu C, Han P, Jin X, Zhang X, He T, et al. HT-Fed-GAN: Federated Generative Model for Decentralized Tabular Data Synthesis. Entropy. 2022;25(1):88. https://doi.org/10.3390/e25010088

61. Zhao Z, Birke R, Kunar A, Chen LY. Fed-TGAN: Federated Learning Framework for Synthesizing Tabular Data. arXiv [Internet]. 2021 [cited 2023 Feb 2]; Available from: http://arxiv.org/abs/2108.07927.

62. Zhao Z, Wu H, Van Moorsel A, Chen LY. GTV: Generating Tabular Data via Vertical Federated Learning. arXiv [Internet]. 2023 [cited 2023 Feb 11]. Available from: http://arxiv.org/abs/2302.01706.

63. Behera MR, Upadhyay S, Shetty S, Priyadarshini S, Patel P, Lee KF. FedSyn: Synthetic Data Generation using Federated Learning. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2203.05931.

64. Dalmaz O, Mirza U, Elmas G, Özbey M, Dar SU, Ceyani E, et al. One Model to Unite Them All: Personalized Federated Learning of Multi-Contrast MRI Synthesis. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2207.06509.

65. Elmas G, Dar SU, Korkmaz Y, Ceyani E, Susam B, Ozbey M, et al. Federated Learning of Generative Image Priors for MRI Reconstruction. IEEE Trans Med Imaging. 2022. https://doi.org/10.1109/TMI.2022.3220757.

66. Fang ML, Dhami DS, Kersting K. DP-CTGAN: Differentially Private Medical Data Generation using CTGANs. In: Michalowski M, Abidi SSR, Abidi S, editors. Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science(), vol 13263. Springer, Cham; 2022. https://doi.org/10.1007/978-3-031-09342-5_17

67. Han T, Nebelung S, Haarburger C, Horst N, Reinartz S, Merhof D, et al. Breaking medical data sharing boundaries by using synthesized radiographs. Sci Adv. 2020;6(49). https://doi.org/10.1126/sciadv.abb7973

68. Li D, Kar A, Ravikumar N, Frangi AF, Fidler S. Federated simulation for medical imaging. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science(), vol 12261. Springer, Cham;2020. p. 159–68. https://doi.org/10.1007/978-3-030-59710-8_16

69. Mei Y, Guo P, Patel VM. Escaping Data Scarcity for High-Resolution Heterogeneous Face Hallucination. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2022. p. 18655–65. https://doi.org/10.1109/CVPR52688.2022.01812

70. Pfitzner B, Arnrich B. DPD-fVAE: Synthetic Data Generation Using Federated Variational Autoencoders With Differentially-Private Decoder. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2211.11591.

71. Rajotte JF, Mukherjee S, Robinson C, Ortiz A, West C, Ferres JML, et al. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. In: Conference on Information Technology for Social Good (GoodIT'21). ACM; 2021. p. 79–84. https://doi.org/10.1145/3462203.3475875

72. Triastcyn A, Faltings B. Federated Generative Privacy. IEEE Intell Syst. 2020;35(4):50–7. https://doi.org/10.1109/MIS.2020.2993966

73. Wang H, Han Y, Yang S, Song A, Zhang T. Privacy-Preserving Federated Generative Adversarial Network for IoT. In: 2021 Int Conf Netw Netw Appl (NaNA). 2021;75–80. https://doi.org/10.1109/NaNA53684.2021.00021

74. Weldon J, Ward T, Brophy E. Generation of Synthetic Electronic Health Records Using a Federated GAN. arXiv [Internet]. 2021 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2109.02543.

75. Welten S, Holt A, Hofmann J, Schelter L, Klopries EM, Wintgens T, et al. Synthetic rainfall data generator development through decentralised model training. J Hydrol. 2022;612:128210. https://doi.org/10.1016/j.jhydrol.2022.128210

76. Xie G, Wang J, Huang Y, Li Y, Zheng Y, Zheng F, et al. FedMed-GAN: Federated Domain Translation on Unsupervised Cross-Modality Brain Image Synthesis. arXiv [Internet]. 2022 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2201.08953.

77. Hayes J, Melis L, Danezis G, De Cristofaro E. LOGAN: Membership inference attacks against generative models. In: Proceedings on Privacy Enhancing Technologies (PoPETs). 2019;1:133–152. Available from: https://petsymposium.org/popets/2019/popets-2019-0008.pdf.

78. Fan S, Xu H, Fu S, Xu M. Smart Ponzi Scheme Detection using Federated Learning. In: 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). 2020. p. 881–8. https://doi.org/10.1109/HPCC-SmartCity-DSS50907.2020.00117

79. Ickin S, Vandikas K, Moradi F, Taghia J, Hu W. Ensemble-based synthetic data synthesis for federated QoE modeling. In: 2020 6th IEEE Conference on Network Softwarization (NetSoft). 2020. p. 72–6. https://doi.org/10.1109/NetSoft48620.2020.9165379

80. Weinger B, Kim J, Sim A, Nakashima M, Moustafa N, Wu KJ. Enhancing IoT anomaly detection performance for federated learning. Digit Commun Networks. 2022;8(3):314–23. https://doi.org/10.1016/j.dcan.2022.02.007

81. Younis R, Fisichella M. FLY-SMOTE: Re-Balancing the Non-IID IoT Edge Devices Data in Federated Learning System. IEEE Access. 2022;10:65092–102. https://doi.org/10.1109/ACCESS.2022.3184309

82. Wang H, Muñoz-González L, Eklund D, Raza S. Non-IID data re-balancing at IoT edge with peer-to-peer federated learning for anomaly detection. In: Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks. 2021. p. 153–63. https://doi.org/10.1145/3448300.3467827

83. Shin M, Hwang C, Kim J, Park J, Bennis M, Kim SL. XOR Mixup: Privacy-Preserving Data Augmentation for One-Shot Federated Learning. arXiv [Internet]. 2020 [cited 2023 Feb 2]. Available from: http://arxiv.org/abs/2006.05148.

84. Taub J, Elliot M, Sakshaug JW. The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records. Trans. Data Priv.. 2020. Apr 19;13(1):1–23. Available from: http://www.tdp.cat/issues16/tdp.a306a18.pdf.

85. Taub J, Elliot M. The Synthetic Data Challenge. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. 2019. Available from: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf.

86. Little C, Elliot M, Allmendinger R, Samani SS. Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. In: 2021 Expert Meeting on Statistical Data Confidentiality. 2021. Available from: https://unece.org/sites/default/files/2021-12/SDC2021_Day2_Little_AD.pdf.

87. Little C, Elliot M, Allmendinger R. Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata. In: Privacy in Statistical Databases PSD 2022. Springer International Publishing; 2022. p. 234–49. https://doi.org/10.1007/978-3-031-13945-1_17

88. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Inf Fusion. 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Abbreviations

AE:	Autoencoder
AUC-ROC:	Area Under Receiver Operating Curve
AUC-PR:	Area Under Precision-Recall Curve
DL:	Deep Learning
DP:	Differential Privacy
EHR:	Electronic Health Records
FedAvg:	Federated Averaging algorithm
FL:	Federated Learning
GAN:	Generative Adversarial Network
iid:	Independent and identically distributed
ML:	Machine Learning
NN:	Neural Network
PRISMA:	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RMSE:	Root mean squared error
SDC:	Statistical Disclosure Control
VAE:	Variational Autoencoder

Supplementary appendices

Please browse Full Text version to see the data of Supplementary Appendix 1.

Supplementary Appendix 1: presents the data charting form with the variables and information extracted from the papers.