AINA 2022, LNNS 450
This is a self-archived pre-print version of this article.
The final publication is available at Springer via
https://doi.org/10.1007/978-3-030-99587-4_27.

Anomaly Detection from Distributed Data Sources via Federated Learning

Florencia Cavallin[1] and Rudolf Mayer[1,2][0000−0003−0424−5999]

[1] SBA Research, Vienna, Austria
[2] Vienna University of Technology, Vienna, Austria
rmayer@sba-research.org

Abstract. Anomaly detection is an important task to identify rare events such as fraud, intrusions, or medical diseases. However, it often needs to be applied on personal or otherwise sensitive data, e.g. business data. This gives rise to concerns regarding the protection of the sensitive data, especially if it is to be analysed by third parties, e.g. in collaborative settings, where data is collected by different entities, but shall be analysed together to benefit from more effective models.
Besides various approaches for e.g. data anonymisation, one approach for privacy-preserving data mining is Federated Learning – especially in settings where data is collected in several distributed locations. A common, global model is obtained by aggregating models trained locally on each data source, while the training data remains at the source. Therefore, data privacy and machine learning can coexist in a decentralised system. While Federated Learning has been studied for several machine learning settings, such as classification, it is still rather unexplored for anomaly detection tasks. As anomalies are rare, they are not picked up easily by a detection method, and the representation in the model dedicated to recognise them might be lost during model aggregation.
In this paper, we thus study anomaly detection task on two different benchmark datasets, in supervised, semi-supervised, and unsupervised settings. We federate Multi-Layer Perceptrons, Gaussian Mixture Models, and Isolation Forests, and compare them to a centralised approach.

Keywords: Federated Machine Learning, Anomaly Detection

1 Introduction

Increasingly, organisations are collecting large volumes of data such as logs, product information, and personal information on clients or customers. The increasing demand for analysing and extracting anomalies, patterns, and possible correlations of these data spurred unprecedented interest in the analysis of this data, propelling some methods to higher effectiveness and efficiency. Alongside, the demand for data sharing and exchange between different parties holding data is increasing, often because different data sets complement each other. Collaborative analysis of data can be beneficial, e.g. to learn from misuse patterns such as fraud that other parties have been exposed to, or in the medical domain.

However, especially when data contains personal information, regulatory, ethical, and security concerns can restrict the potential to fully leverage data, as distribution and exchange are limited. Thus, means to enable collaborative analysis are required. Federated Learning (FL) [1] is a collaborative learning approach that trains models locally across multiple nodes, which each hold their data. This data never leaves the node, and is thus not exposed to the network and possible attacks. The objective of FL is to obtain models with an effectiveness similar as if trained from centralised data. Anomaly detection is an important task in many domains and applications, and users can benefit from exchanging knowledge on their observed anomalies via collaborative learning. However, FL has not yet received much attention in FL research.

In this paper, we thus investigate whether anomaly detection from distributed data via FL can indeed achieve results comparable to a setting where data is centralised. Differences in how unsupervised, semi-supervised, and supervised learning algorithms are affected by federating are investigated. To this end, we analyse the performance of these approaches on two benchmark datasets, from the medical and fraud detection domain. We consider a setting where data is gathered by multiple organisations, and each has a sizeable number of data records. [2] calls this *cross-silo federated learning*, as each of these organisation operates its own data silo. Data is generated locally, and remains decentralised. Regarding the number of clients in cross-silo FL, [2] e.g. talks about 2–100 clients.

We investigate anomaly detection in tasks in health care for detecting diseases, and in identifying fraudulent behaviour in financial transactions such as credit card payments. These two application areas are prototypes for the importance of privacy and confidentiality of the data analysed, as both medical as financial data contain individual data, and are highly sensitive. Further, these two domains are often characterised by individual data silos collecting parts of the overall available data, and hurdles to exchange or centralise it – either due to regulatory, or also due to reservations for sharing sensitive business data. Thus, they are prime candidates for addressing this task in a federated learning setup.

The remainder of this paper is organised as follows. Section 2 discusses related work, before Section 3 describes the federated anomaly detection algorithms we use. Section 4 details the evaluation setup, and Section 5 then discusses the results. Finally, we provide conclusions and future work in Section 6.

2 Related Work

Anomaly detection is the process by which data points, events, and observations that differ from he normal behaviour within a dataset are identified [3]. Although researchers define an anomaly differently based on the application domain, one widely accepted definition is that of Hawkins[4]: 'An anomaly is an observation which deviates so much from other observations as to arouse suspicions that a different mechanism generated it.' Anomalies can point out significant, but rare, events such as technical malfunctions, accidents, or client behaviour changes.

Anomalies may be caused by variations in machine behaviour, fraudulent behaviour, mechanical defects, human error, instrument error and natural deviations in populations [5]. Anomalies in data lead to important actionable information in many application domains, making it a critical task. An unusual traffic pattern in a computer network, for example, could indicate that a hacked computer is transmitting confidential data to an unauthorised recipient. Anomaly detection is employed e.g. in cybersecurity intrusion detection, defect detection of safety-critical devices, health care, fraud detection, or robot behaviour [3].

Anomalies fall into three main categories [3]. A *point anomaly* is an individual data instance that is anomalous with respect to the rest of data. *Contextual anomalies* are data instances that are anomalous in a specific context (of other data, but not otherwise), while a *collective anomaly* denotes a collection of related data instances that are anomalous with respect to the entire data set. In this paper, we address point anomalies.

Anomaly detection can also be distinguished by the availability of labels in the training data [6]. If we consider two types of instances in the data, namely anomalies and normal (regular) data, then we have the following characteristics of training data. In a *supervised* task, we have labels for both anomalies and regular data; this is most often approached with supervised machine learning (classification). If labels are available only for the regular data, we deal with a *semi-supervised* task, where a model is learned for the normal class, and anomalies are those that deviate from that model. If there are no labels available at all, then we deal with a *unsupervised* tasks. In this work, we consider all cases.

The output of an anomaly detection method can be either directly a label (anomaly or normal data), or a score that measures the degree of anomaly, which is then normally compared to a threshold to arrive at a decision.

The privacy and confidentiality of the training data (resp. the individuals represented by it) has been recognised as an important aspect, and thus, privacy-preserving data mining (PPDM) methods are studied. In [7], PPDM techniques are classified into four main categories: (i) data collection privacy, which refers to data randomisation strategies, before they are sent to a data collector, (ii) Privacy-Preserving Data Publishing (PPDP), (iii) Data Mining Output Privacy (DMOP) and (iv) distributed privacy. PPDP often distorts the data, and includes techniques such as k-anonymity, or ϵ-differential privacy. Data synthetisation, which has been studied for various tasks (e.g. [8]), including anomaly detection [9], can also be seen as a form of PPDP. DMOP, on the other hand, which operates on original, unabridged data, relies on ensuring that the computation does not require the exchange of input data. *Federated Learning* can be considered a distributed DMOP method: FL allows to let data remain distributed at the site where it is created, e.g. on mobile devices, respectively where it is initially gathered. However, it still allows to learn a common model from these data, based on aggregating models learned by local training at each site [10]. The idea of local training is relevant for settings where data sharing brings various regulatory, privacy and technical issues, such as the medical domain, or also when sharing business data, e.g. in a collaborative fraud detection setting.

FL is increasingly used in several domains. In [11], the authors showed that federated learning on medical image data can reach a performance, similar as to when data is centralised before training. For structured data, [12] showed that FL is comparable to centralised learning in several settings. In [13], a framework for applying federated learning to biomedical data was presented. In [14] the authors considered Federated Learning for IoT, and optimised the Federated Averaging algorithm of [10] for Edge Computing.

Several forms of collaboration have been investigated in domains that rely on anomaly detection. Collaborative *intrusion detection systems* (CIDSs) [15] address limitations of conventional systems in terms of scalability and massively parallel attacks, CIDSs comprise several monitoring components in a hierarchical structure that collect and exchange data, to eliminate bottlenecks of a centralised approach. [15] identify *privacy* as one of the requirements of a successful collaborative approach – alerts and data exchanged may contain sensitive information that should not be shared. Exchanging only learned knowledge as e.g. in federated learning would be one approach to mitigate these risks.

Exchanging learned knowledge can also be performed by employing transfer learning and domain adaptation, which knowledge learned in one setting is exploited to improve generalisation in another setting [16]. It can leverage information from labelled examples in one domain to predict labels in another domain. This means that models that are useful for one organisation can be transferred to other, similar cases. Transfer learning shows promising results for several task, but for anomaly detection, an open research question is the degree of transferability[17]. The authors of [18] motivate transfer learning as candidate for detection of unknown attack types, and conclude that semi-supervised methods transfer better than supervised ones, but identified a need for improvement. Opposed to FL, transfer learning generally allows only a one-way transfer from a source to a target, and not collaborative learning.

3 Federated Anomaly Detection Algorithms

We use the following algorithms for federated anomaly detection: Multi-Layer Perceptron (for supervised anomaly detection), Gaussian Mixture Model (for semi-supervised anomaly detection) and Isolation Forest (for unsupervised anomaly detection). We describe these and their federated version below.

Multi-Layer Perceptrons (MLPs), a type of *Artificial Neural Network* (ANN), consist of several *neurons* that are arranged in *layers*. Each neuron computes an activation from its inputs and weights. MLPs are feed-forward, i.e. activations are only passed to the next layer, but not backwards. During training, the weights are updated (learned) iteratively, to minimise the error on the training set, by layer-wise back-propagating the gradient of the error and adapting the weights, e.g. via stochastic gradient descent (SGD). If an MLP contains at least two hidden layers, it is a *Deep Neural Network* (DNN) (though the term DNN can denote any ANN with more than one hidden layer, not just MLPs).

It is relatively straightforward to federate an MLP – the *FedAvg* algorithm [10] e.g. performs averaging of the locally trained models. First, a (global) model is initialised, i.e. the weights are randomly set. They are then sent to each client, where they train the model weights further, each with their local training dataset. Subsequently, the clients send their model parameters updates to the central aggregator. FedAvg combines the updates from the clients by averaging, and replaces the previously randomly initialised model with the new weights. This cycle is normally repeated several times, to allow the model to converge. The local training of the same, randomly initialised model at different clients followed by aggregation and averaging was shown to achieve substantially lower loss compared to independently training models on each subset of the data.

In our evaluation, a hyper-parameter optimisation showed that an MLP with two hidden layers achieves best results on the anomaly detection tasks.

Gaussian Mixture Models (GMM) represent a parametric probability density function as a weighted sum of Gaussian *component* densities [19]. They are commonly used for e.g. clustering purposes. GMM parameters (means, μ, and variances, σ^2, of the Gaussians) are estimated from training data using e.g. the Expectation-Maximisation (EM) algorithm by maximum likelihood estimation techniques that maximise the likelihood of a given data sample with the model parameters. Calculating the solution analytically can be mathematically impossible; expectation maximisation is an iterative algorithm and has the property that the maximum likelihood strictly increases with each subsequent iteration, i.e. it is guaranteed to approach a local maximum or saddle point.

Training mixture models does not require having class labels for the data points. A GMM can thus be used in an anomaly detection task when no anomaly cases are known, i.e. in a semi-supervised algorithm. The model then recognises patterns representative of the normal behaviour. When an anomaly sample is to be predicted, the model will likely not group it in any of the identified clusters, since the clusters were created from the normal samples.

We transfer GMMs to federated Gaussian Mixture Models for anomaly detection as follows. First, the the parameters of the global model are randomly initialised for the number of desired Gaussian mixtures (components), and sent to the clients. At each client, the model is trained with the local data, either for a defined number of epochs, or until the certainty of each sample not being part of the assigned cluster is at most a given threshold δ. The averaging to the global model then consists of two steps - finding matching components from each local client, and eventually averaging their parameters.

Isolation Forests (IF) [20]are an unsupervised anomaly detection algorithm. It differs from other approaches, as it is based on *isolating* anomalies, instead of the more common approach of learning a representation of the normal samples. They are based on two assumptions. First, that anomalies are a minority, with very few samples within the dataset. Secondly, that anomalies are different – their values differ notably from normal samples. An Isolation Forest is an ensemble of Isolation Trees, which are binary trees arranging samples by attribute values.

While a Decision Tree, a supervised algorithm, splits data into subsets based on maximising a certain measure (e.g. Information Gain), an Isolation Tree splits based on a random value in the value range of a randomly selected attribute. The number of partitions required to isolate a point is calculated as the length of the path from the root to reach a leaf node. When the Isolation Tree construction is finished, each sample is isolated at a leaf node. Intuitively, anomaly samples are those with a shorter path length in the tree. Based on this, an anomaly score is computed for each instance, and if it is above a predefined threshold (e.g. 0.5), then the sample is labelled as anomaly. The Isolation Forest algorithm has a low linear time complexity, and can be trained with or without anomalies, and in an unsupervised manner [21].

The federated Isolation Forest is implemented as an ensemble of the locally trained Isolation Forests, in a similar manner as in [22].

4 Evaluation Setup

In this section, we describe the setup of our evaluation, including the datasets.

4.1 Datasets

We evaluated our federated anomaly detection algorithms on two benchmark datasets that are frequently used for anomaly detection in centralised settings.

Credit Card Fraud [3] is a dataset that contains 284,807 credit card transactions made by European cardholders over two days in September 2013. Out of the 284,807 transactions, only 492 transactions (0,17%) are fraudulent, making the dataset heavily skewed. The dataset contains 30 features: the amount and time of the transactions, and 28 features obtained via a PCA on the original input data. There are no missing values. Most fraudulent transactions are very small expenditures, probably unnoticed to the cardholders, while the normal samples exhibit all possible values in the range. For preprocessing, the variable "amount" was scaled via a standard scaler to be in line with the other attributes.

Ann-Thyroid [4] is a medical dataset with 3,772 training and 3,428 testing samples, described by 15 categorical and six numerical attributes. There are three possible target values for each instance, namely *normal* (92.583% of the total samples), *hyperfunction* (5.111%) and *subnormal functioning* (2.306%). The hyper function and subnormal classes are treated as the anomaly classes. For supervised detection (with the MLP), where more than one anomaly class can be identified, they will be treated as two separate anomaly classes. For the Gaussian Mixture and Isolation Forest algorithms, these two anomaly classes are merged, in line with other related work. This could influence the effectiveness of the anomaly detection task, as subnormal and hyperthyroid anomalies may have different behaviours.

[3] `https://www.kaggle.com/mlg-ulb/creditcardfraud`

[4] `https://archive.ics.uci.edu/ml/datasets/thyroid+disease`

4.2 Evaluation Metrics

Anomaly detection can be evaluated by several different metrics. As data is normally heavily imbalanced towards the normal class, measures like accuracy (the number of correct predictions) are not sufficient – as already predicting everything to be of the normal class would score very high. Thus, frequently the so-called F-score, a combination of the precision and recall, is employed. *Precision* denotes the ratio of the number of true positive samples (anomalies identified as such) divided by the number of false positives (normal samples wrongly predicted as anomalies). *Recall* is the ratio of samples classified as positive among the total number of positive samples – i.e. how many of the anomalies have been identified. The *F1 score* then provides a combined score by computing the harmonic mean of precision and recall. On the other hand, the *F2 score* weights recall higher than precision. This makes it suitable for the datasets considered in our evaluation, where identifying most of the anomalies is critical, while a certain amount of false positives can be tolerated. However, the actual preference for F1 or F2 (or other F-scores) depends on the exact application scenario, and how many false positive cases can be tolerated and handled, respectively how critical not identified anomalies are, and needs to be determined by domain experts.

Another measure frequently employed is the *area under curve* (AUC) of the receiver operating characteristics (ROC), which is based on the true positive rate (TPR) and false positive rate (FPR), and indicates if we picked randomly a "normal" and anomaly sample, the anomaly example one will have a higher anomaly score, with a probability that corresponds to the AUC. A perfect model will have its AUC equal to 1, while a poor model will have its AUC score around 0. If the AUC is 0.5, it means that the model has no class separation capacity at all.

4.3 Data Distribution

For federated learning, we test different numbers of clients, namely from two to ten with a step size of one, and then 10, 15, 20, 25, 30. In Section 5, due to space limitations we mostly report results for a medium amount of clients, 15, and the largest configuration with 30 clients.

The data is randomly split among the clients to achieve a distributed setting. During our experiments, we use a holdout method to split the data into training and test set in a 90:10 ratio.

5 Results

In this section, we present and discuss our experimental results. We compare the federated results to an idealised, centralised baseline, i.e. where a model can be trained on all data. This represents a glass-ceiling for the federated learning, and is a very difficult baseline to achieve. We thus argue that achieving this glass-ceiling baseline is not necessarily mandated to deem the federated detection as

Table 1: Anomaly detection scores on the credit card fraud dataset

	MLP			GMM			IF		
	Pr	Re	F2	Pr	Re	F2	Pr	Re	F2
Centralized	87.5	85.7	86.1	36.6	39.8	39.1	13.3	4.1	4.7
FL-15 Clients	85.4	71.4	73.8	40.7	52.4	49.6	0.9	16.3	3.8
FL-30 Clients	78.6	44.9	49.1	78.6	44.9	49.1	0.0	0.0	0.0

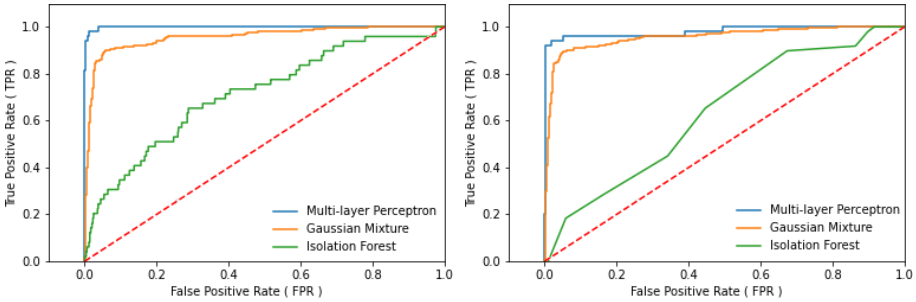
successful, as in many real-world settings, gathering all data in a centralised manner will not be possible.

While we present results from supervised, semi-supervised and unsupervised anomaly detection on the same datasets and with the same algorithms, it has to be noted that these approaches are only partially comparable to each other, and each rather constitute a separate task. This is due to the fact that supervised detection has much more information available when learning the model than the other two approaches (labels for both classes), and is thus an easier task. Unsupervised has the least information available, and is thus the hardest task.

5.1 Credit Card Fraud Dataset

Table 1 shows the scores for the anomaly detection algorithms on the credit card fraud dataset, depicting precision, recall and the F2 score. We can observe that anomaly detection on this dataset is difficult already on the original, centralised setting. The easiest task is the supervised setting, which we address with the Multi-Layer Perceptron (MLP), and which consequently achieves the best scores; its precision, recall and F2 values are all in a very similar range, namely 87.5%, 85.7% and 86.1%, respectively. The Gaussian Mixture Model (GMM), which we employ to solve the semi-supervised task, also has all of its scores within a similar range, albeit lower than the MLP, with values of 36.6% for precision, 39.8% for recall, and 39.1% for F2. Isolation Forests (IF) are used to solve the hardest task, i.e. the unsupervised setting. In line with that difficulty, it scores low on the recall, and thus also achieves a low F2 score. The AUC scores for the centralised setting are shown in Figure 1a, indicating a similar trend.

When comparing the centralised to the federated learning setting, we can notice that the different methods for the tasks are affected to a varying degree. With an increasing number of clients, the MLP loses mostly on recall. Figure 1b shows that also the AUC score drops in the federated setting. This can indicate that the averaging mechanism of FedAvg is not capable of completely preserving the parts of the individual MLPs that learned to represent the anomalies, if the number of clients increases too much, respectively, if there are very few anomaly instances at each client. Strategies to improve this could be e.g. in boosting the weights representing anomalies, similar to the strategy of boosting weights of malicious nodes in the *model replacement* strategy in a federated backdoor attack [23].



(a) Centralised Setting (b) Federated Setting, 15 clients

Fig. 1: ROC Curves for the Credit Card Dataset

Table 2: Anomaly detection scores on thyroid data set

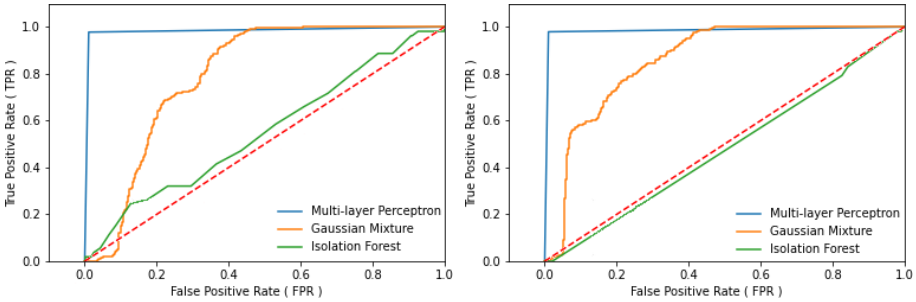
	MLP			GMM			IF		
	Pr	Re	F2	Pr	Re	F2	Pr	Re	F2
Centralized	95.7	83.0	85.3	54.6	37.8	40.3	10.4	18.9	16.2
FL-15 Clients	95.6	81.1	83.7	64.9	62.9	63.3	9.8	14.8	13.4
FL-30 Clients	95.5	79.2	82.0	12.5	1.9	2.2	12.5	1.9	2.2

For GMMs, we can however notice that the F2 score actually increased in the federated setting, while the AUC score stays roughly the same (cf. Figure 1b). The cause for this effect is not systematic – sometimes it is due to a higher recall, but other times due to a higher precision.

For Isolation Forests, the initial trend is similar as for the MLP – a larger number of clients in the federation leads to a drop in effectiveness of the detection, for all scores. A too large number of 30 clients, and thus many Isolation Trees used in the Forest, leads to the anomaly detection not properly working anymore, and no anomalies being detected. This is likely due to the fact that if only few clients have data with anomalies, only a few trees representing these anomalies are created, and they are subsequently outvoted by the many trees representing the normal cases.

5.2 Ann-Thyroid Dataset

Table 2 shows the precision, recall and F2 scores for the anomaly detection algorithms on the thyroid dataset. While the overall best scores for the MLP and GMM on this dataset are comparable to the credit card fraud dataset (with a difference in F2 scores of $\pm$ 1%), the Isolation Forests, albeit on a still low score, performs significantly better on the thyroid dataset. We can observe that for MLP and GMM, precision is significantly better than recall, in the range of 10-15%. With Isolation Forests, that observation is inverse. The AUC scores are depicted in Figure 2a, and indicate a similar trend.



(a) Centralised Setting (b) Federated Setting, 15 clients

Fig. 2: ROC Curve for Thyroid Dataset

When comparing the scores from the centralised to the federated setting, Table 2 and Figure 2b indicate that the trend for the MLP is, albeit dropping, rather stable, i.e. there is a decrease of around 1.5% each when going from centralised to federated learning with 15 clients, and then increasing that number to 30 clients. The drop is mostly due to a drop in recall, as precision stays almost the same for all these settings, just dropping by 0.1% each.

For GMM and Isolation Forests, we can notice that with 30 clients the aggregation into a single model fails – the F2 scores drop to an unusable value, especially due to a low recall. For 15 clients, GMM is however delivering useful results, even better than in the centralised benchmark, mainly due to an increased recall. For Isolation Forests with 15 federated clients, results drop by around 3% from the centralised baseline, but are still significantly better than as for 30 clients. As with the credit card dataset, when too many clients each have only a few anomaly instances, paired with the more difficult semi- and unsupervised tasks, it seem to be a challenge to preserve the knowledge learned on the anomalies during the aggregation process.

6 Conclusions and Future Work

In this paper, we investigated anomaly detection in tasks on medical and financial datasets. We evaluated the performance of three algorithms in central as well as a federated settings, for three settings of availability of labels in the training data – supervised, semi-supervised, and unsupervised. While we observe that especially the supervised method (an MLP) translated very well into the federated setup, the other two methods manged to match their centralise baseline only in some of the settings, especially with a smaller number of federated clients.

We can identify several strands for future work. On the one hand, the adaption of the centralised algorithms into federated versions can be improved, especially for semi- and unsupervised approaches. For Isolation Forest, approaches that e.g. select only the most relevant subset of Isolation Trees might lead to an improvement. Also, further anomaly detection algorithms will be considered.

Further, the centralised baseline where a model is trained on all data is an idealised baseline, and in fact represents a glass-ceiling for the federated learning. It is also an unlikely comparison, as in a real setting, centralising this data is not possible. Another comparison would be to evaluate each locally trained model against the (global) test set; this will simulate how well the anomaly detection works if every client works in isolation, not collaborative, and thus represents a lower bound. To arrive at a realistic judgement on the value of federated anomaly detection, it should be evaluated whether it is well-positioned between these two bounds, and provides a clear advantage over within-silo training.

As studies have shown that unbalanced (in terms of the size of each silo's dataset) and not independent and identically distributed data (non-iid data) can lead to slower convergence, increased communication costs, or lower effectiveness in federated learning [24], we will investigate these for anomaly detection.

Finally, another important aspect to investigate is the vulnerability of federated learning towards inference attacks. Studies such as [25] have shown that federated learning is still vulnerable to e.g. membership inference, and might even open up novel attack vectors. [26] show that outliers might be specifically vulnerable to these attacks, and thus protecting the anomalies is of importance.

Acknowledgments This work was partially funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078 (Project 'FeatureCloud'). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. SBA Research (SBA-K1) is a COMET Centre within the COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

References

1. Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *Workshop on Private Multi-Party Machine Learning, Conf. on Neural Information Processing Systems (NIPS)*, 2016.
2. Peter Kairouz, H. Brendan McMahan, et al. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14(1-2), 2021.
3. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), July 2009.
4. D. M. Hawkins. *Identification of Outliers*. Springer Netherlands, Dordrecht, 1980.
5. Victoria Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), October 2004.
6. Markus Goldstein and Seiichi Uchida. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, 11(4), 2016.
7. Ricardo Mendes and Joao P. Vilela. Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access*, 5, 2017.

8. Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *International Conference on Availability, Reliability and Security*, Canterbury, UK, 2019. ACM.
9. Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. Privacy-Preserving Anomaly Detection Using Synthetic Data. In *Data and Applications Security and Privacy XXXIV*, Cham, 2020. Springer International Publishing.
10. Brendan McMahan, Eider Moore, Daniel Ramage, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 2017. PMLR.
11. Micah J. Sheller, G. Anthony Reina, Brandon Edwards, et al. Multi-institutional Deep Learning Modeling Without Sharing Patient Data. In *BrainLes: International MICCAI Brainlesion Workshop*, Granada, Spain, 2018. Springer.
12. Anastasia Pustozerova, Andreas Rauber, and Rudolf Mayer. Training Effective Neural Networks on Structured Data with Federated Learning. In *Advanced Information Networking and Applications (AINA)*, Cham, 2021. Springer.
13. Santiago Silva, Boris Gutman, Eduardo Romero, Paul M. Thompson, Andre Altmann, Marco Lorenzi, and U K Adni. Federated learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data. Technical report, Inria & Université Cote d'Azur, France, 2018.
14. Jed Mills, Jia Hu, and Geyong Min. Communication-Efficient Federated Learning for Wireless Edge Intelligence in IoT. *IEEE Internet of Things Journal*, 7(7), 2020.
15. Emmanouil Vasilomanolakis, Shankar Karuppayah, Max Mühlhäuser, and Mathias Fischer. Taxonomy and Survey of Collaborative Intrusion Detection. *ACM Computing Surveys*, 47(4), July 2015.
16. Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), December 2016.
17. Raghavendra Chalapathy and Sanjay Chawla. Deep Learning for Anomaly Detection: A Survey, January 2019. arXiv: 1901.03407.
18. Chuanliang Chen, Yunchao Gong, and Yingjie Tian. Semi-supervised learning methods for network intrusion detection. In *International Conference on Systems, Man and Cybernetics*, Singapore, Singapore, October 2008. IEEE.
19. Douglas Reynolds. Gaussian Mixture Models. In *Encyclopedia of Biometrics*. Springer US, Boston, MA, 2009.
20. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, December 2008. IEEE.
21. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1), March 2012.
22. Yang Liu, Yingting Liu, Zhijie Liu, Yuxuan Liang, Chuishi Meng, Junbo Zhang, and Yu Zheng. Federated Forest. *IEEE Transactions on Big Data*, 2020.
23. Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How To Backdoor Federated Learning. In *23rd Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Italy, 2020. PMLR.
24. Felix Sattler, Simon Wiedemann, Klaus-Robert Muller, and Wojciech Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), September 2020.
25. Anastassiya Pustozerova and Rudolf Mayer. Information Leaks in Federated Learning. In *Proceedings 2020 Workshop on Decentralized IoT Systems and Security*, San Diego, CA, 2020. Internet Society.
26. Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning*. PMLR, 2021.