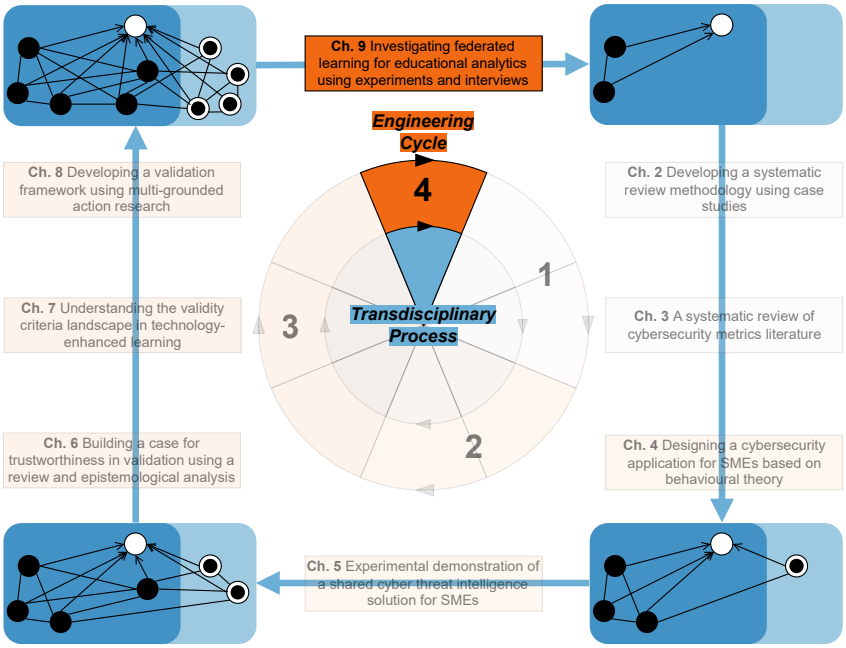
Transdisciplinary perspectives on validity: bridging the gap between design and implementation for technology-enhanced learning systems

Haastrecht, M.A.N. van

Part IV

TREATMENT IMPLEMENTATION



9

FEDERATED LEARNING FOR EDUCATIONAL ANALYTICS

Concerns surrounding privacy and data protection are a primary contributor to the hesitation of institutions to adopt new educational technologies. Addressing these concerns could open the door to accelerated impact, but current state-of-the-art approaches centred around machine learning are heavily dependent on (personal) data. Privacy-preserving machine learning, in the form of federated learning, could offer a solution. However, federated learning has not been investigated in-depth within the context of educational analytics, and it is therefore unclear what its impact on model performance is. In this chapter, we compare performance across three different machine learning architectures (local learning, federated learning, and central learning) for three distinct prediction use cases (learning outcome, question correctness, and dropout). We find that federated learning consistently achieves comparable performance to central learning, but also that local learning remains competitive up to 20 local clients. We introduce FLAME, a novel metric that assists policymakers in their assessment of the privacy-performance trade-off, and conclude by discussing preliminary findings from a series of interviews with stakeholders we are conducting to unearth their views on federated learning for education.

The contents of this chapter are based on: van Haastrecht, M. Brinkhuis, and Spruit (2024). Federated Learning Analytics: Investigating the Privacy-Performance Trade-off in Machine Learning for Educational Analytics. Accepted at AIED 2024, prior to inclusion of interview materials in discussion.

9.1 INTRODUCTION

Driven by the promise of analytics to enable learning environment optimisation, education is now more datafied than ever (Williamson et al., 2020). The large-scale collection of learner data raises concerns regarding ethics, privacy, fairness, and trustworthiness (Gardner et al., 2023; van Haastrecht, M. Brinkhuis, Peichl, et al., 2023). Research tends to focus on the data protection measures educational institutions should implement to convince learners that they can be trusted as data fiduciaries (Jones et al., 2020). Examples of suggested measures concerning data that has already been collected are limiting the boundaries of access to student data, pseudonymisation and anonymisation of learner records, and using automated bias mitigation. However, approaches that assume that personal data has already been collected fail to address a fundamental question: Did we have to collect the data in the first place?

It is not trivial to motivate which, if any, educational optimisations would warrant an intrusion of student privacy. Institutes that hold student privacy in high regard may be of the opinion that collecting personal learning data is never warranted (Rubel and Jones, 2016). This puts educational analytics research in an uncomfortable position, as methods and applications commonly rely heavily on personal data. Machine learning models such as deep neural networks predicting learning outcomes (Waheed et al., 2020) and transformers facilitating student knowledge tracing (D. Shin et al., 2021) are deeply dependent on the availability of large amounts of data. On the surface, it seems as though these data-hungry machine learning models are incompatible with a policy of preserving student privacy. However, in recent years we have seen the development of machine learning architectures that promise the performance of machine learning without the threats to privacy posed by institute access to personal data.

Privacy-preserving machine learning architectures such as federated learning (McMahan et al., 2017), where only model parameters are shared with a centrally coordinating party, offer a promising future direction for educational analytics. Along with local learning, where nothing is shared, and central learning, where everything is shared, federated learning is among the major machine learning architectures to consider from a privacy perspective. We have recently seen the first studies investigating the promise of federated learning for educational analytics (Fachola et al., 2023; Guo and Zeng, 2020). However, to our knowledge, no study has systematically compared local learning, federated learning, and central learning across different datasets and use cases. This is a significant gap in the literature when we consider that privacy-preserving techniques could be the key to giving control back to students (Ekuban and Domingue, 2023).

In this chapter, we hope to take a first step in systematically investigating the promise of federated learning for learning analytics, which we term 'federated

learning analytics'. We compare the performance of local learning, federated learning, and central learning across three distinct use cases: learning outcome prediction, question correctness prediction, and dropout prediction. Our methodology is geared at answering our main research question:

- **RQ**: How does the privacy-performance trade-off for machine learning algorithms manifest itself in different educational analytics use cases?

9.2 BACKGROUND

Preserving the privacy of learners while actively collecting their data has long been recognised as a major challenge. It is evident that students should never be considered simply as sources of data, but rather as collaborators whose learning and development we are trying to serve (Slade and Prinsloo, 2013). However, although the importance of formulating and employing ethical and privacy principles was recognised early on, privacy concerns regularly played second fiddle due to the "enthusiasm for the possibilities offered by learning analytics" (Prinsloo and Slade, 2015). New legislation surrounding data protection introduced new perspectives. Besides ethical and privacy concerns, legal concerns began to drive decisions made at educational institutions. In the educational privacy framework DELICATE (Drachsler and Greller, 2016), the section on legitimacy contains the question: "Which data sources do you have already, and are they not enough?" Questions like these represented a major change of mindset. Researchers and practitioners recognised that collecting particular types of data is never warranted, and that "learning analytics is justifiable just to the extent that it does indeed promote autonomy" (Rubel and Jones, 2016).

Basic organisational and technical controls can help to preserve student privacy, but it is questionable whether this is sufficient to gain students' trust. Prinsloo and Slade (2015) convincingly argue that "the power to harvest, analyse and exploit data lies completely with the provider," rather than the student. The authors outline the importance of transparency towards students and of giving students the possibility to access and update their own information. However, the issue with these measures is that they still require the student to entrust multiple stakeholders with their personal data, keeping alive the privacy power imbalance between the student and the data fiduciary.

Levelling out the power balance is exactly what decentralised approaches have attempted to do in recent years, by enabling the sharing of student data in a way that can enhance both privacy and security within educational systems. Students thus regain some ownership over their data, helping to restore the power balance. Yet, using a decentralised architecture also introduces new challenges. The most prominent of these is how to maintain performant algorithms when not all data is available in one central data store. A study of several anonymisation and differential privacy techniques found that in

a GPA prediction task accuracy could drop from 76% to anywhere between 45-63% (Gursoy et al., 2017). Novel methods such as deep learning and transformers are notorious for requiring immense datasets to tune their parameters. How can we continue using these successful machine learning architectures when we do not have the data they so desperately need in one central location?

McMahan et al. (2017) introduced the concept of federated learning, where learning occurs over a federation of users referred to as clients. Rather than having to share data and parameters, clients train their model on local data and only share the parameter values of their model with the coordinating server. By averaging the parameters of all local clients, the resulting global model obtains better performance than if all local clients operated independently. Figure 9.1 visualises the scenarios of local learning, federated learning, and central learning. A fourth scenario was recently proposed where data is kept locally and parameters are not shared with a centrally coordinating server, but rather with other trusted parties via blockchain (Warnat-Herresthal et al., 2021). This architecture, termed swarm learning, is worth considering for educational institutions. However, we will not investigate it in detail within this chapter.

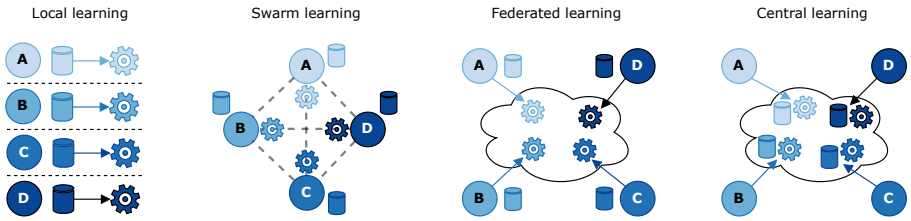


Figure 9.1: Visualisation of various machine learning architectures (based on (Warnat-Herresthal et al., 2021)). In the local learning scenario both data and parameters remain at the client. Federated learning only shares model parameters, whereas swarm learning removes the need for a centrally coordinating server and shares model parameters over blockchain while keeping data at client nodes. For central learning, both data and parameters are shared with a centrally coordinating server.

Decentralised machine learning could be the key towards privacy-preserving, trustworthy educational analytics (Ekuban and Domingue, 2023). Yet, only a couple of studies have investigated this promising area. Guo and Zeng (2020) use federated learning in the context of educational data analysis. They consider the task of dropout prediction in the KDD Cup 2015 dataset, achieving accuracy within a couple of percentage points of the central learning scenario. However, the authors do not make their code available and do not report performance metrics other than a single figure showing accuracy progression over epochs. This concern about their work was voiced by a more recent fed-

erated learning paper using the KDD Cup 2015 dataset. Fachola et al. (2023) achieve an accuracy of 81.7% in the case of central learning and show that using federated learning an accuracy of around 80% can be achieved, even when data is spread over more than 50 clients. A downside is that the reported accuracy of 81.7% is only two percentage points higher than the proportion of dropouts in the dataset of 79.3%. Accuracy is not the right choice of metric for this dataset. If we want to draw meaningful conclusions about the potential of federated learning analytics, we need to consider multiple datasets and performance metrics.

9.3 METHODOLOGY

This section describes the metrics we used to compare the performance of different models, the three datasets (OULAD, EdNet, and KDD Cup 2015) employed in our experiments, and the details of our federated learning algorithm.

9.3.1 *Metrics*

Two commonly used metrics to evaluate model performance are accuracy and F_1 score. Accuracy represents the fraction of correctly predicted records. The F_1 score is the harmonic mean of precision p (true positives divided by all predicted positives) and recall r (true positives divided by all actual positives). Both metrics should be used with caution when dealing with imbalanced datasets, as they are influenced heavily by whether the majority class is appointed as the positive or negative class.

A metric that is less explicitly sensitive to class imbalance is the Area Under the ROC Curve (AUC). The curve in question is a plot of the true positive rate (equal to recall) on the y-axis and the false positive rate (false positives divided by all actual negatives) on the x-axis. The curve is drawn by determining the true positive rate and the false positive rate at different classification thresholds, meaning AUC requires the probability estimates of a model for its calculation. Because AUC is based on probability outputs, rather than the 0-1 classification output, it can provide more fine-grained insight into whether a model is truly learning to separate positive from negative instances. AUC does suffer from its own issues, such as that it can be biased towards certain classifiers.

9.3.2 *Datasets*

The Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017), contains demographic data on students and logs of student activity within a virtual learning environment. The outcome variable of interest is

Table 9.1: Descriptive statistics of the three datasets we investigate in this chapter: OULAD, EdNet, and KDD Cup 2015. We additionally indicate state-of-the-art (SOTA) results for each, where the OULAD metrics are divided into PF (pass-fail), PW (pass-withdrawn), FD (fail-distinction), and PD (pass-distinction).

	OULAD (Waheed et al., 2020)	EDNET (D. Shin et al., 2021)	KDD CUP 2015 (W. Feng et al., 2019)
Use case	learning outcome	question correctness	dropout
# Students	32,593	784,309	200,902
# Records	10,655,280	95,293,926	13,545,124
% Pos. class	PF: 31% fail PW: 40% withdrawn FD: 30% distinction PD: 20% distinction	66% correct	79% dropout
SOTA	PF: Acc.=0.845 F_1=0.719 PW: Acc.=0.947 F_1=0.943 FD: Acc.=0.864 F_1=0.770 PD: Acc.=0.805 F_1=0.749	Acc.=0.725 AUC=0.791	F_1=0.929 AUC=0.909

the result a student achieved for a course, which can be pass, distinction, fail, or withdrawal. OULAD forms the basis for studies varying from the creation of predictive models identifying at-risk students (Hlosta et al., 2017) to the investigation of the role of demographics in virtual learning environments (Rizvi et al., 2019). We use the work of Waheed et al. (Waheed et al., 2020) as our baseline for comparison, as the authors provide a detailed description of the features they use, allowing us to conduct a replication that closely matches their process. They turn the original classification problem with four potential outcomes into four separate binary classification tasks (pass=0 & fail=1, pass=0 & withdrawn=1, fail=0 & distinction=1, pass=0 & distinction=1). Table 9.1 reports the accuracy and F_1 score achieved for each of these tasks.

EdNet is a knowledge tracing dataset containing data from users of a self-study platform (Choi, Y. Lee, D. Shin, et al., 2020). Rather than having a single outcome variable per user, EdNet involves predicting for each completed multiple-choice question whether a user answered it correctly. The prediction task of EdNet is temporal in nature, explaining why papers tackling this dataset tend to employ time-series machine learning models such as transformers (Choi, Y. Lee, J. Cho, et al., 2020). We use the SAINT+ transformer model (D. Shin et al., 2021) as our baseline for comparison, as this is the model with the current state-of-the-art performance. The authors use a version of EdNet with newer user data that is not publicly available. Yet, since the prediction task and features are identical, their results can still serve as a useful benchmark.

The final dataset we consider was used for the KDD Cup 2015 challenge. This dataset contains information on student interactions within a Massive Open Online Course (MOOC) environment. The goal is to predict student dropout, with a distinguishing characteristic being that 79% of the enrolled students dropped out. The dataset is thus highly imbalanced, explaining why KDD Cup 2015 papers tend to focus on reporting AUC and F_1 scores, rather than accuracy (W. Feng et al., 2019; W. Li et al., 2016).

9.3.3 *Federated learning*

Federated learning was proposed as a communication-efficient way to use all available data on individual devices to train a global model, without users having to share their personal data (McMahan et al., 2017). The use case considered when introducing swarm learning was that of a group of hospitals working together to create better predictive models for the detection of illnesses (Warnat-Herresthal et al., 2021) . The sensitivity of health data, along with the extensive legislation limiting data sharing in medical settings, provides a clear motivation for the need for a parameter-sharing infrastructure without a centrally coordinating party. A recent study in the educational field investigated a transfer learning approach and voiced concerns regarding the relevance of decentralised approaches for education (Gardner et al., 2023). Hence, we should ask to what extent decentralised machine learning contexts appear in educational environments.

Guo and Zeng (2020) and Fachola et al. (2023) envision a network of schools that are part of a federation sharing model parameters. These schools are part of the same governing body, but have separate physical locations, possibly even in different countries. From a legal and privacy perspective, it can then be worthwhile to employ federated learning to obtain optimal insight into student behaviour without needing to share student data across schools. The use case considered in both papers is dropout prediction using the KDD Cup 2015 dataset, meaning each student has a single outcome variable per course. Federated learning on the level of the classroom or the individual is likely not realistic here, since the majority of students have fewer than five course outcomes to train on. For the KDD Cup 2015 dataset we will therefore investigate federated learning performance up to a maximum of 100 local clients, corresponding to roughly 2,000 students per client. OULAD is comparable to the KDD Cup 2015 dataset, with the exception that it additionally contains demographic information. For OULAD we similarly analyse up to 100 local clients, corresponding to roughly 300 students per client.

For the EdNet setting, where a single student can answer thousands of questions in their self-study process, federated learning with individual students as local clients is more realistic. Nevertheless, since single users potentially have only one answered question within EdNet, it is not algorithmically practical to have local clients comprising one user. In our experiments, we

Table 9.2: Comparison of our central learning results to the results of Table 9.1, where the value between brackets represents the performance difference with earlier work.

	OULAD		EdNet		KDD Cup 2015	
	ACC.	F_1	ACC.	AUC	F_1	AUC
PF	0.862 (+0.017)	0.751 (+0.032)	0.720 (-0.005)	0.757 (-0.035)	0.925 (-0.003)	0.881 (-0.028)
PW	0.933 (-0.014)	0.914 (-0.011)				
FD	0.893 (+0.029)	0.820 (+0.050)				
PD	0.810 (+0.005)	0.199 (-0.551)				

will investigate the performance of local and federated learning up to a maximum of 100 local clients, corresponding to around 100 users per client when working with a randomly selected subset of 10,000 students.

9.4 RESULTS

The Python code used to produce the outcomes of this section and detailed results per dataset are available on GitHub[1]. Our federated learning code adheres to the `FedAvg` algorithm of McMahan et al. (2017). Central learning experiments were conducted using the machine learning library scikit-learn and the gradient boosting libraries XGBoost and CatBoost. We used Pytorch as the deep learning library for our federated learning algorithm and exclusively used XGBoost with default settings as our local learning classifier.

9.4.1 *Central learning*

Table 9.2 presents our central learning results using 10-fold cross-validation with an 80-20 train-test split. Our best results were achieved using CatBoost (OULAD and KDD Cup 2015) and XGBoost (EdNet). Table 9.2 shows that we managed to achieve comparable performance to the current state-of-the-art.

Since Waheed et al. (2020) extensively describe the features they engineered, we were able to reproduce these features and use them as input for OULAD classification. For the EdNet prediction task, we created lag features for previous user question correctness to turn the time series prediction task into a classification task. This enabled us to utilise the regular machine learning and gradient boosting libraries we used for OULAD and KDD Cup 2015. For the KDD Cup 2015 dataset, we designed student activity features similar to those of OULAD.

1 https://github.com/MaxvanHaastrecht/Federated-Learning-Analytics

9.4.2 *Local learning and federated learning*

For our local and federated learning scenarios, we divided students randomly over clients. For OULAD federated learning, we used a neural network with two hidden layers of sizes 30 and 10, a learning rate η of 0.02, a cross-entropy loss function with the Adam optimiser, the number of communication rounds R set to 50, the number of local epochs per round $E = 2$, and a batch size of 64. Figure 9.2 shows that both federated learning and local learning perform worse than the central learning scenario. However, whereas local learning accuracy drops significantly as we progress from 10 to 100 local clients, federated learning accuracy remains roughly constant.

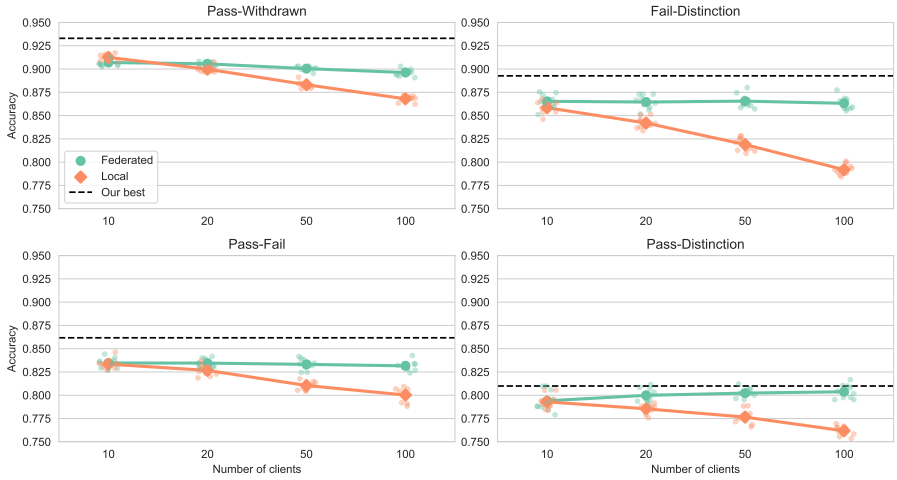


Figure 9.2: Plot of the bootstrapped mean accuracy for varying numbers of local clients, showing comparisons of our local learning, federated learning, and central learning results.

Figure 9.3 summarises the results from our EdNet and KDD Cup 2015 experiments. For KDD Cup 2015, we used the exact same federated learning settings as with OULAD. For EdNet, we changed the batch size to 128, as is used in earlier work (Choi, Y. Lee, J. Cho, et al., 2020), and lowered the number of communication rounds R from 50 to 20. We additionally used hidden layer sizes of 16 and 8, rather than 30 and 10, since EdNet feature engineering resulted in fewer input features for the network. Since the EdNet dataset is comparatively large, it is common practice to work with a random subset of the dataset in experimental settings such as our federated learning context (Long et al., 2022; Y. Yang et al., 2021). We work with a random subset of 10,000 students and indicate the AUC of our best central learning model in Figure 9.3.

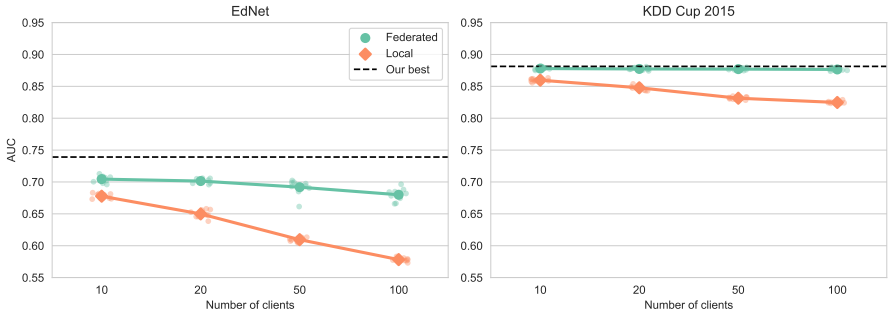


Figure 9.3: Plot of the bootstrapped mean AUC for varying numbers of local clients, showing comparisons of our central learning EdNet and KDD Cup 2015 AUC results to local learning and federated learning.

9.4.3 *Federated learning analytics metric (FLAME)*

Our numerical results provide an indication of the performance of federated learning compared to local learning and central learning. However, our results are not directly usable by policymakers in education deciding whether to opt for a federated learning architecture. Questions remain regarding the optimal number of local clients in each scenario and how much performance we are willing to trade off for an improved preservation of privacy. To ease the decision-making process, we propose the federated learning analytics metric (FLAME). The idea behind FLAME is to capture the trade-off between privacy and performance in a single metric, such that comparisons across scenarios, datasets, and numbers of local clients become more tenable. We define FLAME as:

$$\text{FLAME} = \frac{1 - \frac{1}{K}}{1 + (p_c - p_f)} = \frac{\text{privacy gain}}{1 + \text{performance loss}},$$

where K is the number of local clients, p_c is the central learning performance, and p_f is the federated learning performance. For institutions considering to move from a central learning architecture to federated learning, p_c will be a known quantity. For institutions that do not have a centralised architecture, p_c can be estimated based on the literature or through simulations. FLAME is suited to be used for performance metrics ranging between [0,1], such as accuracy, F_1, and AUC. The numerator captures the gain in privacy achieved by employing an architecture with local clients. The denominator captures the loss in performance.

Figure 9.4 shows the FLAME values for EdNet and KDD Cup 2015, where AUC is the relevant performance metric. FLAME values for the local learning scenario are also shown, which can be calculated by replacing the federated learning performance in the FLAME formula with local learning performance.

Taking EdNet as an example, we observe that for federated learning FLAME peaks at 50 clients, whereas for local learning FLAME peaks at 20 clients. By more explicitly incorporating the privacy-performance trade-off, FLAME therefore clarifies differences between algorithms in a way the pure AUC scores of Figure 9.3 cannot.

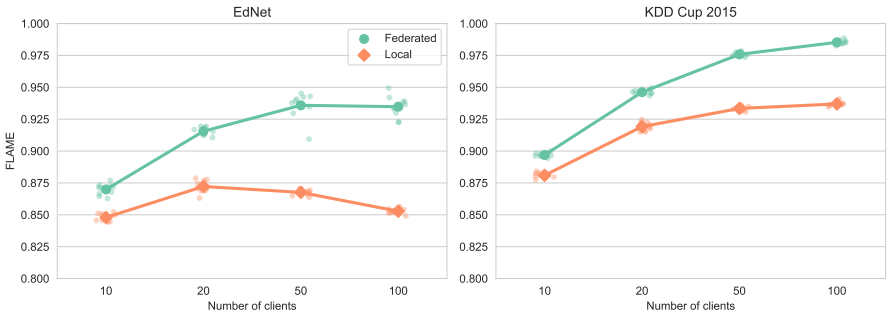


Figure 9.4: FLAME values for EdNet and KDD Cup 2015, where AUC is the performance metric. In the case of 50 local clients, AUC loss must be less than 0.0315 to achieve a FLAME higher than 0.95.

9.5 DISCUSSION

Our results demonstrate the potential of federated learning to preserve privacy and performance in educational contexts. For OULAD, we observed that our federated learning algorithm achieved comparable accuracy to earlier results for three out of four scenarios considered, even when the number of local clients was set to 100. For the KDD Cup 2015 dataset, federated learning matched our best results, again up to 100 local clients. Federated learning also significantly outperformed local learning for all three datasets. When dividing data over 100 local clients, the average accuracy gain for OULAD was 4.32% and the average AUC gains for EdNet and KDD Cup 2015 were 0.1017 and 0.0518, respectively.

Our FLAME values in Figure 9.4 demonstrated that local learning and federated learning warrant serious consideration in settings where dividing data over 20 or more clients is realistic. However, the answer to student privacy concerns can never be purely technological. Federated learning is promising, but it carries with it additional security risks and questions whether student's perceptions of these technologies are as positive as their theoretical benefits. Yet, given the increasing tensions between the datafication of education and the privacy concerns of students, privacy-preserving machine learning architectures may offer the path of least resistance towards a bright future for educational analytics.

Federated learning is perhaps the most commonly used privacy-preserving machine learning strategy, but certainly not the only one. We did not cover other paradigms within this chapter, such as split learning (Thapa et al., 2022), swarm learning (Warnat-Herresthal et al., 2021), and transfer learning (Gardner et al., 2023). In future work, it will be crucial to compare the privacy-performance trade-off for various approaches. We should be aware that in contexts where performance takes precedent, combining strategies (e.g., federated learning and split learning (Thapa et al., 2022)) might be the optimal choice, whereas in contexts where privacy is paramount, a local learning approach that fosters stakeholder trust could provide the perfect fit. Regardless of the privacy-preserving paradigms considered, insights regarding the privacy-performance trade-off provided by FLAME can serve as a useful starting point for discussion.

A limitation of our work is that all benchmarking datasets had drawbacks. OULAD is extensively documented and publicly available, but is comprised of scenarios with imbalanced classification tasks where the metrics currently used in the literature (accuracy and F_1) are inadequate for thorough comparisons of model performance. EdNet is publicly available, but recent work has relied on a version of the dataset that is not publicly available (D. Shin et al., 2021), or has worked with subsets of the full dataset that hinder replicability (Long et al., 2022; Y. Yang et al., 2021). The KDD Cup 2015 dataset is not publicly available from a dedicated website, and the most relevant publications covering this dataset in recent years only report model accuracy (Fachola et al., 2023; Guo and Zeng, 2020), when this is a highly imbalanced dataset with 79% of students dropping out. These drawbacks are not ideal, but we strongly believe these datasets offer an accurate representation of currently available benchmarks. Still, we require better benchmark datasets and accompanying research in the future.

9.5.1 *Interviews with stakeholders*

To uncover the views of stakeholders at educational institutions regarding federated learning, we plan to conduct a follow-up study where we use a grounded theory approach to analyse the data resulting from a series of qualitative interviews. The analysis of the first two interviews with educational technology experts in higher education have been completed at this stage, and we deem it relevant to report two preliminary findings here.

Firstly, the experts we interviewed pointed out that federated learning could serve as a stepping stone for educational institutions to move from experimental situations to wide-scale impact. Interestingly, the two experts both used the metaphor of a chicken and egg situation, whereby a prerequisite to scale up an educational innovation is a demonstration of its impact, but to demonstrate impact you need the data of students that you only get after you scale up. One of the interviewees put it as follows: "It's kind of chicken

and egg situation. To demonstrate that an algorithm can be trusted, you have to do some kind of analysis, so people can see that it offers advantages and makes education better. But then you have to be able to start, and if there is suspicion regarding an innovation then you can never start anything." Federated learning could help to perform the required analysis without immediately having to implement a solution that is not trusted by students.

Secondly, one of the interviewees explained why they consider it worthwhile to keep developing privacy-preserving machine learning techniques with regards to the concept of proportionality: "If you have no other option than central learning, then you can talk all you want about proportionality, but then you have no choice. If you can use different methods, you can try to find a balance in privacy risk and usability." In other words, if we do not keep developing privacy-preserving machine learning techniques for education, cases will occur where our only realistic option is central learning. We will then find ourselves in a situation where we cannot adequately attend to the proportionality principle which is central to regulations such as GDPR.

9.6 CONCLUSION AND FUTURE WORK

With education becoming more datafied than ever, researchers interested in optimising learning environments are increasingly faced with questions regarding ethics, privacy, fairness, and trustworthiness. Decisions to intrude on student privacy should be taken with the utmost caution. There are legitimate concerns whether any type of optimisation warrants the collection of sensitive learner data. Within this context, privacy-preserving machine learning that respects privacy while maintaining model performance is an intriguing recent development. However, until now, we lacked rigorous investigations of the impact of privacy-preserving architectures on educational analytics model performance.

We compared algorithm performance across three architectures (local learning, federated learning, central learning) for three different prediction use cases (learning outcome, question correctness, dropout). In doing so, we provided a comprehensive image of what can be achieved with privacy-preserving architectures. We found that even when dividing data over 100 clients, federated learning can compete with state-of-the-art results. A major finding was that although for 50 or more clients federated learning outperformed local learning, differences were often not significant when dividing data over 20 or fewer clients. This points to the importance of considering local learning as a privacy-preserving strategy for educational analytics. Future work will need to extend the investigation of how students, teachers, and other stakeholders view federated learning, since the relative complexity of privacy-preserving machine learning may diminish trust. Nevertheless, as evidenced by the preliminary findings from our interviews with stakeholders,

the datafication of education combined with the clear wish of students to preserve privacy signal a promising future for federated learning analytics.