Privacy-Preserving Knowledge Transfer with Bootstrap Aggregation of Teacher Ensembles

Hong-Jun Yoon[1](✉), Hilda B. Klasky[1], Eric B. Durbin[2], Xiao-Cheng Wu[3], Antoinette Stroup[4], Jennifer Doherty[5], Linda Coyle[6], Lynne Penberthy[7], Christopher Stanley[1], J. Blair Christian[1], and Georgia D. Tourassi[8]

[1] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
{yoonh,klaskyhb,stanleycb,christianjb}@ornl.gov

[2] College of Medicine, University of Kentucky, Lexington, KY 40536, USA
ericd@kcr.uky.edu

[3] Louisiana Tumor Registry, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA
XWu@lsuhsc.edu

[4] New Jersey State Cancer Registry, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901, USA
nan.stroup@rutgers.edu

[5] Utah Cancer Registry, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84132, USA
Jen.Doherty@hci.utah.edu

[6] Information Management Services Inc., Calverton, MD 20705, USA
coylel@imsweb.com

[7] Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD 20814, USA
lynnepenberthy.schumacher-penberthy@nih.gov

[8] National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
tourassig@ornl.gov

Abstract. There is a need to transfer knowledge among institutions and organizations to save effort in annotation and labeling or in enhancing task performance. However, knowledge transfer is difficult because of restrictions that are in place to ensure data security and privacy. Institutions are not allowed to exchange data or perform any activity that may expose personal information. With the leverage of a differential privacy

This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

V. Gadepally et al. (Eds.): Poly 2020/DMAH 2020, LNCS 12633, pp. 87–99, 2021.
https://doi.org/10.1007/978-3-030-71055-2_9

algorithm in a high-performance computing environment, we propose a new training protocol, Bootstrap Aggregation of Teacher Ensembles (BATE), which is applicable to various types of machine learning models. The BATE algorithm is based on and provides enhancements to the PATE algorithm, maintaining competitive task performance scores on complex datasets with underrepresented class labels.

We conducted a proof-of-the-concept study of the information extraction from cancer pathology report data from four cancer registries and performed comparisons between four scenarios: no collaboration, no privacy-preserving collaboration, the PATE algorithm, and the proposed BATE algorithm. The results showed that the BATE algorithm maintained competitive macro-averaged F1 scores, demonstrating that the suggested algorithm is an effective yet privacy-preserving method for machine learning and deep learning solutions.

Keywords: Data privacy · Privacy-preserving machine learning · Differential privacy · Bootstrap aggregation · Information extraction · Natural language processing

1 Introduction

Data security and privacy are prime topics in the design of artificial intelligence (AI) systems [14,15,17,18]. Domains such as biomedical and health informatics, finance, tax revenue services, and homeland security characteristically use sensitive data that contains personal information about human subjects. For the safety of the data and personal information, exchanging such data among organizations and institutions is strictly controlled to prevent any possible leakage of sensitive human subject information. However, to develop faithful deep learning (DL)-based machine learning (ML) information extraction and classification models, a large amount of data from various data sources is highly desirable. Moreover, some institutions may be limited by having too few training examples to achieve ML/DL models to meet their expectations [10]. Thus, there is a need for ways to transfer knowledge securely among organizations and institutions.

However, current AI and ML-based data processing approaches present security vulnerabilities that can be exploited to leak sensitive details. Exposure of private information can occur as a result of the features captured by DL models. A key feature of DL models is that they equip multiple layers of trainable parameters. They learn by example and extract optimal feature representations to enable higher accuracy. However, the ML/DL training algorithm is domain-agnostic and does not recognize if a feature contains sensitive information.

Privacy-preserving models aim to prevent the (identification and) storage of sensitive information in training data used in ML algorithms. Privacy herein is understood as establishing a differential privacy approach that identifies privacy with a measurable and rigorous mathematical definition [5]. Differential privacy allows companies to collect the data of users without compromising the privacy of the individuals [7]. Differential privacy [6] ensures that the probability distribution of the released statistics is roughly similar without paying attention to the inclusion or non-inclusion of any single member in the study; thus, it provides credible statistics. Applying differential privacy to ML algorithms provides a very strong guarantee that the datasets can be shared across registries without concerns about privacy and confidentiality.

The study described in this document focuses on the application and evaluation of an approach based on the Private Aggregation of Teacher Ensembles (PATE) algorithm [16]. PATE is a modified teacher/student model that includes differential privacy [16]. PATE initiates by working on a set of sensitive data, which is partitioned into different sections that do not overlap. On each partition, any ML model, which in PATE's framework is called a "teacher," is trained. The set of ML models or teachers is called an "assemble." Different and independent learning models can be used in each partition separately. At the inference phase of the algorithm, PATE aggregates the predictions of the teacher assemble. To do so, PATE counts votes, adds Laplace noise to the teachers' answers, and then takes the maximum value. The Laplace noise introduces randomness to protect the privacy of users when the teachers do not have a strong quorum. The final step is to transfer the knowledge from the teacher assemble model to a student model using some public data (unlabeled). The teacher assemble will label some of the unlabeled public data, and the student model will contain a training set that will be used to learn a model and perform predictions. The student model is added to decrease the probability of total privacy loss. In recent years, there have been several refinements to the PATE model; specifically, there have been improvements to the student model part of the algorithm [19].

One limitation that we observed with PATE is that it is too conservative regarding underrepresented and minor classes during the classification process. To address this issue, we propose an enhancement to the PATE framework. We named our approach BATE (Bootstrap Aggregation of Teacher Ensemble). BATE uses bagging (bootstrap aggregating) in high-performance computing (HPC), instead of the data partitioning implemented in the PATE framework. Thus, it yields performance scores for the minor classes at the same time that it ensures differential privacy. We hypothesize that including the bootstrap classification will help improve BATE HPC performance.

In this paper, we performed a feasibility study of the proposed BATE model with the data from four cancer registries. We developed models to extract information on morphological and topographical characteristics of tumors from cancer pathology reports. The cancer pathology dataset was labeled by the cancer/tumor/case codes that met the Surveillance, Epidemiology, and End Results (SEER) case reporting guidelines. We developed multitask convolutional neural network (MT-CNN) models and confirmed that the model was feasible for the cancer pathology report corpus [4]. In this study, we simulated a scenario in which one cancer registry had no gold standard labels and so learned from the other three registries. But there was a restriction that no cancer registry should expose patients' identities and information to others. The results presented in this study support our hypothesis on performance improvement. We show that the performance of the BATE model is superior to that of the PATE model, especially for subsite and histology, those classes suffering from severe class imbalance, and many underrepresented classes.

This paper is organized as follows: Sect. 2 presents related work, and Sect. 3 presents the data and methods used in the study. Section 4 presents the results. Section 5 provides a discussion, limitations, conclusions, and future work.

2 Related Work

Part of the groundwork that established the foundation of PATE was the work on differential privacy on neural networks by Abadi et al. [2]. That study introduced a differential privacy stochastic gradient descent (SGD) algorithm aimed at controlling the influence of the training data stage, specifically in SGD computation. In a subsequent study, PATE was presented by Papernot et al. [16] as an independent approach to a learning algorithm for either teacher or student models, i.e. a black-box approach, and consequently, capable of being applied to other learning methods. PATE improved the accuracy of a private MNIST model from 97% to 98% and the privacy bound from 8 to 1.9 [16]. Note that MNIST is a simple classification task. The following are other variations of the PATE approach:

- A PATE variation called PATE-G was introduced by Abadi et al. [3]. PATE-G implements generative methods based on generative adversarial networks (GANs) and semi-supervised models for knowledge transfer, thus improving accuracy and privacy.
- PATE-GAN [11] uses GAN's capabilities to generate synthetic data based on real data using a modified PATE, allowing it to tightly bound the influence of any individual sample on the model. This approach results in tight differential privacy guarantees and thus improved performance over models with the same guarantees.
- In Papernot et al. [19], PATE is applied to larger-scale learning tasks and real-world datasets. Aggregators were improved to allow the application of PATE to uncurated data; in addition, Laplace noise was replaced with Gaussian noise.
- G-PATE [13] also leverages GANs to produce synthetic datasets with strong privacy guarantee. G-PATE ensures differential privacy in the student generator.
- TrPATE [21] modified the original PATE framework and adopted transfer learning to alleviate PATE's performance degradation problem.

However, none of those approaches were applied to bioclinical data, and we found only a limited number of studies applying PATE to bioclinical data. An example is the work of Fay et al. [8,9]. Their studies applied PATE variations to brain tumor segmentation magnetic resonance imaging, as shown in the following references:

- To reduce the required noise level during the aggregation stage, Fay et al. [8] assessed principal component analysis for dimensionality reduction to map the prediction target onto a low-dimensional latent space theoretically and auto-encoders experimentally on a brain tumor dataset. Their study used Gaussian noise in the aggregation stage.
- Autoencoder-based PATE [9] is a PATE variant that builds low-dimensional representations of segmentation masks that the student can obtain through low-sensitivity queries to the private aggregator. This approach achieves a higher Dice coefficient (segmentation quality) for the same privacy guarantee on a brain tumor segmentation dataset.

To our knowledge, at the time of this study, there are no published studies of PATE or any of the PATE variants that address performance scores for the minor classes that have also been applied to bioclinical data on high-performance computers. To help solve these issues, we present the BATE approach, which uses bagging in HPC instead of the data partitioning implemented in the PATE framework. Thus, it generates performance scores for the minor classes at the same time that it includes differential privacy.

The main contributions of our work to the problem we are exploring are the following: 1. We proposed the use of BATE to enhance the PATE differential privacy approach with the use of bagging. 2. We applied BATE to bioclinical data. The results of our study show improvements in those classes suffering from severe class imbalance. 3. The study was performed on a high-performance computer.

In the following section we present the datasets employed in this study and the implementation approach.

3 Methods

3.1 Datasets

The dataset for this study consists of unstructured text in pathology reports from four cancer registries: the Louisiana Tumor Registry, Kentucky Cancer Registry, Utah Cancer Registry, and New Jersey State Cancer Registry. These registries contribute to the National Cancer Institute's SEER program. The study was executed in accordance with the institutional review board protocol DOE000152.

Certified tumor registrars manually coded the ground truth labels associated with each unique case based on free text from the corresponding pathology reports, according to the SEER program coding and staging manual. We consulted the International Classification of Diseases for Oncology, Third Edition, coding convention for labeling the cases. We extracted the following six data fields from the cancer reports: cancer site (70 classes), subsite (320 classes), laterality (7 classes), histology (571 classes), behavior (4 classes), and tumor grade (9 classes). Note that the dataset has a severe class imbalance among class labels (e.g., C50: 242,427 cases, C39: 6 cases), and some labels have few training samples available (e.g., C630: 2 cases, C764: 3 cases), mainly because of the low prevalence of rare cancers.

We chose reports with specimens collected in or after 2017 as testing data and specimens collected in or before 2016 as training data. We randomly selected and reserved 10% of the training data for validation of the model training. Also, we considered only cases for there was less than a 10-day difference between the date of diagnosis and either the specimen collection date or the date of surgery. The 10-day time difference was determined based on an analysis of the pathology report submissions. The vast majority of reports and addenda fell within that time frame. Table 1 lists the number of pathology reports from the four registries. Note that we renamed the registries in the table for security purposes. Note also that, in each registry, there are around 50,000 words in the vocabulary that appeared across the registries.

In this paper, we designed a study in which we selected one registry as a student institution and developed an information extraction DL model with the training data from the other three registries regarding teacher institutions. We repeated this training procedure four times, once per each registry as a student.

Table 1. Number of training and testing cases from the four cancer registries, number of vocabularies in the corpus, and the number of unique words only appearing in the registry. We renamed the registries for security purposes.

Cancer registry	1	2	3	4
Train	147,191	91,820	243,475	259,699
Test	1,554	21,411	58,049	49,433
# Words	479,570	79,959	189,037	247,555
# Unique Words	428,957	35,332	125,787	187,079

3.2 Multi-task Convolutional Neural Networks

We chose the MT-CNN [4] as our DL model for information extraction from cancer pathology reports. It is an extension of the CNN for sentence classification [12,20]. The model consists of three parts: word embedding, one-dimensional convolution, and a task-specific, fully connected layer.

Word embedding is a learned representation of terms to map a set of words onto vectors of numerical values that have the same semantic meaning and have similar observations. A security vulnerability in the word embedding layer is that we can hypothesize that the disease types and the patients' personal information may be clustered together in the vector space.

The convolution layer has a series of one-dimensional convolution filters that have latent representations to capture the features from the word vectors of documents. The algorithm determines the optimal features by itself. However, in the overfitting instances, feature learners may attempt to extract personal identities and sensitive information and become vulnerable to purposeful adversarial attacks.

3.3 Bootstrap Aggregation of Teacher Ensembles Algorithm

In AI and ML, there are two types of information the ML models can observe from the data corpus. One is "common information" that contains concepts and ideas that can articulate data characteristics and their class association. The other is "private information" that is specific to the individual cases and typically should not be contributed to the classification and ML process. However, in certain circumstances, some pieces of private information can be included in the decision process; we refer to the inclusion of such data as "overfitting." The main idea of the PATE algorithm is to divide one training corpus into several subsets that are disjointed from one another and to develop multiple teacher models. The choice of disjoint sets prevents private information from influencing the decision, thus preventing exposure of the identities of individual data subjects in the sensitive data. However, the disjoint data splitting in PATE may cause a considerable performance decrease for decisions in underrepresented classes.

Bootstrap Aggregation. We propose to apply bagging, which is the technique that we do training with multiple models with many sampled data with replacement, thus improving stability and accuracy while helping to avoid overfitting of the data. Both (disjoint) sampling and bagging prevent the extraction of private information from the data; the latter approach maintains or improves performance in classifying minor classes. One drawback is that bagging increases the computational demands for training many models with multiples of the data [22].

Additive Laplace Noise. We added Laplace noise to the teachers' aggregation of predictions, perturbing the counts, and formed a single prediction, which is also known as the "noisymax mechanism" [6]. The purpose of this procedure is to prevent a single outlier from driving decisions when two output classes receive an equal number of votes from the teachers, which could result in the exposure of private information. Additive random noise will not change the decision if it is obvious, thus receives majority votes. Adding a larger scale of noise to the decisions might increase the privacy budget, but it would considerably degrade the overall task performance.

Student Model. Even if the teacher models were trained in a privacy-preserving manner, releasing the models directly to other parties and institutions would present a potential risk of leaking private information because there is a finite privacy budget in the model. Moreover, in cases of natural language processing models, exposure of vocabularies may give a hint to an adversary. Instead, a student institution provides a pilot dataset, and the teacher models derive decisions from the dataset. The student institutions develop their own models based on the pilot dataset with the teacher's annotations.

3.4 Study Design

We designed a study with four participating cancer registries, simulating a situation in which each registry learns from the other three registries. We set up four scenarios as follows:

Scenario 1: No Collaboration. There was no interaction or communication among the cancer registries. Each registry developed its own DL model based on its data and manual annotation. This was the most secure and privacy-preserving method of development, but each institution had to spend effort on it.

Scenario 2: No Privacy-Preserving Collaboration. One institution received a model developed by the data collected from the other three institutions. There was no preparation for an adversary attack on privacy or leaking of personal identity. An institution did not have to spend effort to develop its models, and there was the possibility of a performance boost to some extent because of the abundance of training samples from other institutes.

Scenario 3: PATE. We followed the PATE algorithm: we made 20 disjoint subsamples from the training dataset collected from the three institutes, developed 20 DL models (teacher), and developed one student model trained by the pilot dataset and annotations from 20 teachers. A considerable performance decrease was expected, especially on subsite and histology classification tasks, because those tasks contained several underrepresented class labels.

Scenario 4: BATE. We trained 200 bootstrap sampled datasets and developed 200 teachers. The student model was trained by the pilot dataset with annotations from the 200 teachers. This was the most computationally expensive method of all the scenarios. For the PATE and BATE algorithms in this study, we chose one cancer registry as a student institution and regarded the training set of the registry as the pilot dataset. The student model was trained not by the gold standard of the training set, but by the teachers.

4 Results

We ran experiments in extracting information from cancer pathology reports provided by the four SEER cancer registries, based on the four scenarios described in the previous sections. We extracted the following six properties: primary cancer site, subsite, laterality, histology, behavior, and grade. We performed parallel training and validation of the DL models on the Summit supercomputer operated by the Oak Ridge Leadership Computing Facility (OLCF). The codes were implemented with the Keras and TensorFlow [1] backend available in the IBM Watson ML packages. Since the datasets had many class labels and some had severe class imbalances, we adopted micro and macro-averaged F1 scores as performance metrics. The results are listed in Table 2.

Table 2. Information extraction task performance in micro and macro-averaged F1 metrics for each registry as a student institution and the average from all four registries. S1 (Scenario 1): no collaboration, S2 (Scenario 2): no privacy preservation, S3 (Scenario 3): PATE algorithm, and S4 (Scenario 4): BATE algorithm. λ is the scale of the additive Laplace noise to the aggregated decisions from the teachers.

	Reg.	S1	S2	S3			S4		
λ				0	0.05	0.1	0	0.05	0.1
Site									
Micro F1	1	0.9344	0.9286	0.9279	0.9292	0.9241	0.9324	0.9356	0.9228
	2	0.9287	0.9278	0.9232	0.9218	0.9203	0.9254	0.9264	0.9225
	3	0.9247	0.9202	0.9187	0.9171	0.9130	0.9225	0.9229	0.9215
	4	0.9248	0.9222	0.9145	0.9144	0.9093	0.9238	0.9215	0.9186
	Average	0.9281	0.9247	0.9211	0.9206	0.9167	0.9260	0.9266	0.9213
Macro F1	1	0.6173	0.6491	0.6186	0.6201	0.5892	0.5959	0.6020	0.5934
	2	0.6244	0.6473	0.5653	0.5641	0.5508	0.6209	0.6257	0.6024
	3	0.6424	0.6704	0.5645	0.5544	0.5190	0.6338	0.6373	0.6163
	4	0.6545	0.6355	0.5473	0.5456	0.5080	0.6374	0.6417	0.6102
	Average	**0.6346**	**0.6506**	**0.5739**	**0.5710**	**0.5418**	**0.6220**	**0.6267**	**0.6056**
Subsite									
Micro F1	1	0.5978	0.5927	0.5882	0.5759	0.5592	0.6004	0.6010	0.5766
	2	0.6578	0.6513	0.6439	0.6431	0.6355	0.6634	0.6637	0.6492
	3	0.6435	0.6347	0.6153	0.6135	0.6003	0.6530	0.6543	0.6444
	4	0.6467	0.6490	0.6310	0.6302	0.6231	0.6548	0.6531	0.6429
	Average	**0.6365**	**0.6319**	**0.6196**	**0.6157**	**0.6045**	**0.6429**	**0.6430**	**0.6283**
Macro F1	1	0.3794	0.3573	0.3094	0.2963	0.2907	0.3170	0.3207	0.3068
	2	0.3087	0.3143	0.2391	0.2357	0.2107	0.2953	0.2975	0.2596
	3	0.2771	0.2956	0.2148	0.2037	0.1849	0.2870	0.2818	0.2515
	4	0.3060	0.3029	0.2203	0.2127	0.1874	0.2940	0.2891	0.2574
	Average	**0.3178**	**0.3175**	**0.2459**	**0.2371**	**0.2184**	**0.2983**	**0.2973**	**0.2688**
Laterality									
Micro F1	1	0.9157	0.9028	0.9125	0.9138	0.9151	0.9208	0.9176	0.9118
	2	0.9130	0.9021	0.9029	0.9029	0.8977	0.9030	0.9038	0.9006
	3	0.9036	0.9003	0.9048	0.9054	0.9030	0.9048	0.9041	0.9007
	4	0.9023	0.8982	0.9012	0.9005	0.8990	0.9045	0.9042	0.8983
	Average	**0.9086**	**0.9009**	**0.9053**	**0.9056**	**0.9037**	**0.9083**	**0.9074**	**0.9029**
Macro F1	1	0.5920	0.4693	0.5536	0.5544	0.5592	0.4777	0.4726	0.5092
	2	0.5221	0.5179	0.5097	0.5155	0.5057	0.5201	0.5295	0.5116
	3	0.5296	0.5124	0.5004	0.5066	0.4915	0.5210	0.5231	0.5047
	4	0.5265	0.5141	0.5086	0.5095	0.5039	0.5167	0.5173	0.5039
	Average	**0.5426**	**0.5034**	**0.5180**	**0.5215**	**0.5151**	**0.5089**	**0.5106**	**0.5074**
Histology									
Micro F1	1	0.7252	0.7207	0.7130	0.7072	0.6963	0.7291	0.7246	0.7079
	2	0.7469	0.7414	0.7310	0.7304	0.7217	0.7411	0.7394	0.7300

(*continued*)

Table 2. (*continued*)

	Reg.	S1	S2	S3			S4		
λ				0	0.05	0.1	0	0.05	0.1
	3	0.7546	0.7453	0.7469	0.7462	0.7425	0.7607	0.7613	0.7536
	4	0.7803	0.7756	0.7674	0.7639	0.7580	0.7755	0.7783	0.7674
	Average	**0.7518**	**0.7457**	**0.7396**	**0.7369**	**0.7296**	**0.7516**	**0.7509**	**0.7397**
Macro F1	1	0.4004	0.3906	0.3047	0.3224	0.2544	0.3609	0.3732	0.3472
	2	0.3540	0.3444	0.2245	0.2193	0.1873	0.3096	0.3120	0.2605
	3	0.3239	0.3164	0.1998	0.1859	0.1532	0.3009	0.2981	0.2517
	4	0.3275	0.3276	0.2041	0.1993	0.1452	0.3142	0.3036	0.2551
	Average	**0.3515**	**0.3448**	**0.2333**	**0.2317**	**0.1850**	**0.3214**	**0.3217**	**0.2786**
Behavior									
Micro F1	1	0.9704	0.9743	0.9698	0.9665	0.9659	0.9736	0.9710	0.9646
	2	0.9654	0.9585	0.9595	0.9575	0.9560	0.9598	0.9617	0.9570
	3	0.9671	0.9665	0.9684	0.9678	0.9644	0.9696	0.9688	0.9672
	4	0.9731	0.9670	0.9693	0.9680	0.9655	0.9709	0.9703	0.9680
	Average	**0.9690**	**0.9666**	**0.9667**	**0.9650**	**0.9630**	**0.9685**	**0.9680**	**0.9642**
Macro F1	1	0.8159	0.8595	0.7201	0.7533	0.6460	0.8730	0.8076	0.8373
	2	0.9038	0.8664	0.8094	0.8511	0.8073	0.8554	0.8654	0.8507
	3	0.8133	0.8316	0.7378	0.7393	0.7057	0.8363	0.8151	0.7581
	4	0.8207	0.8389	0.7674	0.7542	0.7265	0.8417	0.8267	0.7893
	Average	**0.8384**	**0.8491**	**0.7587**	**0.7745**	**0.7214**	**0.8516**	**0.8287**	**0.8089**
Grade									
Micro F1	1	0.7259	0.6731	0.6763	0.6692	0.6744	0.6737	0.6660	0.6577
	2	0.7807	0.7732	0.7728	0.7651	0.7673	0.7817	0.7801	0.7752
	3	0.7255	0.7115	0.7196	0.7210	0.7112	0.7279	0.7293	0.7204
	4	0.7587	0.7461	0.7564	0.7552	0.7498	0.7571	0.7586	0.7503
	Average	**0.7477**	**0.7260**	**0.7313**	**0.7276**	**0.7257**	**0.7351**	**0.7335**	**0.7259**
Macro F1	1	0.7364	0.7090	0.7055	0.7046	0.6989	0.7200	0.6910	0.6871
	2	0.6011	0.6297	0.5885	0.5812	0.5803	0.6067	0.6327	0.5968
	3	0.6503	0.6083	0.5631	0.5697	0.5558	0.6220	0.6269	0.5771
	4	0.7716	0.6961	0.6720	0.6728	0.6664	0.6772	0.7453	0.6703
	Average	**0.6898**	**0.6608**	**0.6323**	**0.6321**	**0.6253**	**0.6565**	**0.6740**	**0.6328**

We observed that the F1 scores between S1 (no collaboration) and S2 (transfer knowledge without privacy preservation) were very close. That finding was the confirmation that we could achieve a similar level of task performance by developing a model with the other registries' data and testing it with the student registry. It implies that our study design is legitimate.

Based on the comparisons of F1 scores between S2 and S3 or S4, we observed a certain level of performance decrease if we applied privacy-preserving algorithms; that finding is confirmed by other studies [2] showing that there is a trade-off between accuracy and privacy. However, we observed more degradation of performance from applying the PATE algorithm (S3) than from applying the BATE (S4). That was especially true for the macro-averaged F1 scores of subsite (S3 averaged macro-F1: 0.2459, S4: 0.2983) and histology (S3: 0.2333, S4: 0.3214), two tasks that have many underrepresented class labels. It was a clear demonstration that BATE performance was superior to PATE performance.

The results also supported the findings of other studies that adding more noise to the decision may increase the privacy budget but decrease the classification accuracy [16]. Both the S3 and S4 scenarios showed that increasing the scale parameter of the additive Laplace noise lowered the classification accuracy scores. Those findings were more clear for the macro-F1 scores of the subsite and histology tasks. Also, we observed that BATE performance was superior to PATE performance for the subsite (S3: 0.2184, S4: 0.2688) and histology (S3: 0.1850, S4: 0.2786) tasks.

5 Discussion

Threats to data privacy in AI and ML/DL are incurred as a result of the nature of the design: ML/DL models are trained without having domain knowledge but find the best feature representations that can maximize the task performance. During the training, the algorithm may learn too precisely from the examples and may attempt to extract personal and sensitive information. The state-of-art differential privacy algorithms are designed primarily to avoid such incidents so that the few marginal training samples dominate decisions. We suggested the BATE algorithm, in which we adopted the advantages of PATE so that we could isolate the vocabulary sets of the student institutions from the teacher institutions and limit the access of teacher models to secure privacy. Also, with the BATE model, we maintained the accuracy scores of the underrepresented classes of the training samples.

Information extraction from cancer pathology reports was our model example. There were many class labels in the dataset, including those of rare cancer types, which resulted in severe class imbalances and underrepresentation of training examples. We demonstrated that BATE performance was superior to PATE performance, especially for those difficult problems. We also showed that, with BATE, the privacy-preserving training and transfer of knowledge from the teacher institutions to the student institutions maintained the clinical task performance.

The study's limitation is that we examined the effects of the BATE algorithm qualitatively but did not quantify the threat of privacy and security attacks from the adversary. The results suggested that we need to design a follow-up study to confirm the validity and security of the privacy-preserving knowledge transfer.

Acknowledgement. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the US Department of Energy (DOE) Office of Science and the National Nuclear Security Administration. This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by DOE and the National Cancer Institute of the

National Institutes of Health. This work was performed under the auspices of DOE by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and ORNL under Contract DE-AC05-00OR22725.

KCR data were collected with funding from NCI Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201800013I), the CDC National Program of Cancer Registries (NPCR) (U58DP00003907) and the Commonwealth of Kentucky.

LTR data were collected using funding from NCI and the Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201800007I), the CDC's National Program of Cancer Registries (NPCR) (NU58DP006332-02-00) as well as the State of Louisiana.

NJSCR data were collected using funding from NCI and the Surveillance, Epidemiology and End Results (SEER) Program (HHSN261201300021I), the CDC's National Program of Cancer Registries (NPCR) (NU58DP006279-02-00) as well as the State of New Jersey and the Rutgers Cancer Institute of New Jersey.

The Utah Cancer Registry is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201800016I, and the US Centers for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP0063200, with additional support from the University of Utah and Huntsman Cancer Foundation.

The study was supported by the Laboratory Directed Research and Development (LDRD) program of Oak Ridge National Laboratory, under LDRD project No. 9831.

This research used resources of the Oak Ridge Leadership Computing Facility at ORNL, which is supported by the DOE Office of Science under Contract No. DE-AC05-00OR22725.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 2016), pp. 265–283 (2016)
2. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
3. Abadi, M., et al.: On the protection of private information in machine learning systems: two recent approches. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pp. 1–6. IEEE (2017)
4. Alawad, M., et al.: Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J. Am. Med. Inform. Assoc. **27**(1), 89–98 (2020)
5. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
6. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
7. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends® Theoret. Comput. Sci. **9**(3–4), 211–407 (2014)

8. Fay, D., Sjölund, J., Oechtering, T.J.: Private learning for high-dimensional targets with pate (2020)
9. Fay, D., Sjölund, J., Oechtering, T.J.: Decentralized differentially private segmentation with pate. arXiv preprint arXiv:2004.06567 (2020)
10. Fung, B.C., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. (CSUR) **42**(4), 1–53 (2010)
11. Jordon, J., Yoon, J., van der Schaar, M.: Pate-GAN: generating synthetic data with differential privacy guarantees (2018)
12. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
13. Long, Y., Lin, S., Yang, Z., Gunter, C.A., Li, B.: Scalable differentially private generative student model via pate. arXiv preprint arXiv:1906.09338 (2019)
14. McMahan, H.B., et al.: A general approach to adding differential privacy to iterative training procedures. arXiv preprint arXiv:1812.06210 (2018)
15. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., Talwar, K.: Machine learning with privacy by knowledge aggregation and transfer
16. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755 (2016)
17. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814 (2016)
18. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P.: Sok: security and privacy in machine learning. In: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 399–414. IEEE (2018)
19. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú.: Scalable private learning with pate. arXiv preprint arXiv:1802.08908 (2018)
20. Qiu, J.X., Yoon, H.J., Fearn, P.A., Tourassi, G.D.: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J. Biomed. Health Inform. **22**(1), 244–251 (2017)
21. Wang, L., Zheng, J., Cao, Y., Wang, H.: Enhance pate on complex tasks with knowledge transferred from non-private data. IEEE Access **7**, 50081–50094 (2019)
22. Yoon, H.J., et al.: Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports- manuscript submitted for publication