This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

Privacy-Preserving Deep Learning NLP Models for Cancer Registries

Mohammed Alawad, Hong-Jun Yoon, Shang Gao, Brent Mumphrey, Xiao-Cheng Wu, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Linda Coyle, Lynne Penberthy, and Georgia Tourassi

Abstract—Population cancer registries can benefit from Deep Learning (DL) to automatically extract cancer characteristics from the high volume of unstructured pathology text reports they process annually. The success of DL to tackle this and other real-world problems is proportional to the availability of large labeled datasets for model training. Although collaboration among cancer registries is essential to fully exploit the promise of DL, privacy and confidentiality concerns are main obstacles for data sharing across cancer registries. Moreover, DL for natural language processing (NLP) requires sharing a vocabulary dictionary for the embedding layer which may contain patient identifiers. Thus, even distributing the trained models across cancer registries causes a privacy violation issue. In this paper, we propose DL NLP model distribution via privacy-preserving transfer learning approaches without sharing sensitive data. These approaches are used to distribute a multitask convolutional neural network (MT-CNN) NLP model among cancer registries. The model is trained to extract six key cancer characteristics – tumor site, subsite, laterality, behavior, histology, and grade – from cancer pathology reports. Using 410,064 pathology documents from two cancer registries, we compare our proposed approach to conventional transfer learning without privacy-preserving, single-registry models, and a model trained on centrally hosted data. The results show that transfer learning approaches including data sharing and model distribution outperform significantly the single-registry model. In addition, the best performing privacy-preserving model distribution approach achieves statistically indistinguishable average micro- and macro-F1 scores across all extraction tasks (0.823,0.580) as compared to the centralized model (0.827,0.585).

Index Terms—Privacy-preserving, multi-task CNN, transfer learning, NLP, information extraction, cancer pathology reports.

1 Introduction

Accurate, timely, and comprehensive cancer monitoring is critical for not only assessing the population level impact of cancer but also for informing population-based cancer control policies. Population cancer registries process annually large volumes of unstructured pathology reports to extract cancer characteristics such as tumor anatomic location site, histological type, tumor grade, and stage at diagnosis for reporting to the national cancer surveillance programs. Such critical information resides in narrative text full of typos, abbreviations, and linguistic variation. Natural language processing (NLP) has been explored extensively in oncology to semi-automate the time-consuming and laborious manual effort [1], [2]. Scalable NLP can have a dramatic impact in cancer surveillance by assisting cancer registries in providing near real time detailed measurements of cancer incidence, progression, survival, and mortality. However, existing clinical NLP methods are mainly rule-based requiring human experts to manually engineer input features. This is an unsustainable endeavor due to the prohibitively large number of rules that need to be carefully curated by domain experts to comprehensively capture all possible linguistic expressions. Therefore, artificial intelligence (AI) could potentially address clinical NLP challenges [3] and facilitate effective translation of NLP tools across cancer registries.

Among different AI approaches, Deep Learning (DL) has been successfully applied to classify and recognize complex features in images, speech, and text data. Recent studies have shown the potential of DL models in automatically extracting cancer key characteristics from cancer pathology reports [4], [5], [6], [7] by achieving accuracy superior to traditional machine learning NLP methods. Successfully applying DL in the specific domain requires a large training corpus that has similar characteristics as the prospective testing data. Furthermore, this success is proportional to the size of the training corpus. Obtaining a large enough corpus from a single cancer registry is challenging, particularly

- *M. Alawad, HJ Yoon, S. Gao, and G. Tourassi are with the Computational Sciences and Engineering Division, Health Data Sciences Institute, Oak Ridge National Laboratory, Oak Ridge, TN, 37831. E-mail: {alawadmm,tourassig}@ornl.gov*
- *B. Mumphrey and X-C Wu are with Louisiana Tumor Registry, Louisiana State University Health Sciences Center School of Public Health, New Orleans, LA, 70112.*
- *E.B. Durbin, I. Hands and D. Rust are with Kentucky Cancer Registry, University of Kentucky, Lexington, KY, 40506.*
- *E.B. Durbin and J.C. Jeong are with Division of Biomedical Informatics, College of Medicine, University of Kentucky, Lexington, KY, 40506.*
- *E.B. Durbin, J.C. Jeong, and I. Hands are with Cancer Research Informatics Shared Resource Facility, Markey Cancer Center, University of Kentucky, Lexington, KY, 40506.*
- *L. Coyle is with the Information Management Services Inc, Calverton, MD, 20705.*
- *L. Penberthy is with the Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA.*

Manuscript received Month XX, YEAR; revised Month XX, YEAR.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paidup, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy 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).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

with respect to rare cancer anatomic location sites (i.e., body organs where cancer develops) and histologies (i.e., different cell types). This challenge can be overcome by aggregating cancer pathology reports from multiple cancer registries in a centralized hub which can serve as a neutral entity to train a generalized model on all the data. Upon completion of training, the trained model can be shared with the registries. However, data privacy and confidentiality concerns prevent cancer registries from sharing patient data and benefiting from each other's knowledge by leveraging DL.

Transfer learning can be exploited to avoid data sharing by distributing learning models across cancer registries instead of distributing pathology reports. In transfer learning, a model can be developed at one clinical site, and then reused as a starting point at another clinical site. Therefore, a cancer registry can benefit from other registries labeled datasets to get a more generalized model and reach better performance by using fewer training samples on its end. Although the transfer learning approach has been widely and successfully used in many computer vision applications [8], applying the same approach on text applications and sharing the whole model across data holders still requires access to the source data dictionary which includes sensitive information, such as patient names and residential addresses. Without a universally accepted de-identification algorithm, large scale de-identification is not currently a viable option across cancer registries. Image-based DL models do not contain any individually-identifiable patient information; however, text-based DL models contain such information as part of the word embeddings. To distribute a trained text-based DL model across cancer registries, the vocabulary dictionary, which contains individually-identifiable patient information, must be distributed too. Therefore distributing DL NLP models across cancer registries poses privacy concerns.

This work builds upon our previous work [9], in which we implemented a conventional transfer learning (TL) approach among cancer registries and applied it on a single task CNN model for cancer subsite extraction from pathology reports. We also compared the model trained via TL with a model trained on centrally hosted data. The main contributions of this work are as follows:

- We develop a multitask CNN (MT-CNN) model for information extraction from cancer pathology reports. It differs from the previous work [7] by extracting information at the pathology report level instead of the tumor-level. Also, we consider all available classes of cancer characteristics without condensing low prevalent classes. The model is used to extract six key cancer characteristics – tumor anatomic location site (i.e., site) (70 classes), subsite (313 classes), laterality (7 classes), behavior (4 classes), histology (543 classes), and grade (9 classes).
- We propose a new privacy-preserving approach that protects any PHI information in the word embedding vocabulary dictionary by applying restrictions on which word tokens are included in the vocabulary. To prevent PHI information such as patient names and residential addresses from being included, we limit the vocabulary to words from publicly available corpus that has been prescreened for PHI, such as the MIMIC-III dataset and the PubMed abstracts dataset. Thus, a trained model can be shared with other registries without data restrictions.
- We evaluate the effectiveness of collaboration across cancer registries on the performance of the MT-CNN using different TL methods with and without our privacy-preserving vocabulary. These methods are necessary in scenarios where cancer registries are unable to directly share their patient data for training. We compare the conventional TL approach, acyclic TL, and the state-of-the-art model distribution approach, cyclic transfer learning [10]. Cyclic transfer learning has been used in medical imaging applications, but to our knowledge this is the first time it is applied to medical text. We compare these approaches against the baselines of training the MT-CNN on data from only a single registry and training the MT-CNN on data from all available registries without any restrictions.

2 Related Work

Collaboration among cancer registries through sharing raw data or trained models is hindered by security and privacy violations. One approach of data collaboration without privacy violation is through text de-identification by detecting and scrubbing protected health information (PHI) including name, social security number, geographic identifiers, and dates from cancer pathology reports; the sanitized data can then be shared with other institutes for research purposes. Since manual de-identification approach is costly and time consuming, different techniques have been proposed to support automatic clinical text de-identification using traditional machine learning [11] and DL models [12]. However, automatically locating and scrubbing all sensitive information from clinical text is still highly challenging – de-identifying unstructured text in pathology reports is more challenging than structured data [13], and de-identification models trained on a specific dataset do not generalize well to other datasets [14]. Existing solutions typically cannot guarantee de-identification up to regulatory standards, especially with scattered PHI across the unstructured text of pathology reports; therefore, there is still need for alternative privacy-preserving methods to protect PHI and directly or indirectly share large corpora of pathology reports from various sources.

Another approach of secure collaboration among cancer registries is through DL model distribution. Parallelizing the training of deep networks by distributing the process across computing nodes has been proposed to handle large datasets and accelerate DL models training. Recently, this approach has inspired the effort of privacy-preserving DL. It protects confidential features by distributing a trained model without sharing raw data. Sharing the raw data across clinical institutes can be protected by sharing the trained models. However, to achieve a highly protected mechanism, model characteristics including the architecture, parameters and loss function have to be protected as well. Some of these techniques, such as federated learning [15] and large batch synchronous stochastic gradient de-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

scent SGD [16], require sharing the model hyperparameters, parameters and intermediate representations without any protection. Hitaj et al. [17] have shown the ability of generative adversarial networks to recover raw data from the shared model. Other techniques like SplitNN [18] protect the model parameters; however, they require a relatively larger overall communication bandwidth [19]. The challenge of model distribution increases when dealing with DL NLP models. Such models may include personally identifiable information as part of the word embeddings – each word in the dataset vocabulary is represented by an N-dimensional vector. To distribute a trained model across different institutes, the vocabulary list with word embeddings must be distributed as well, which may contain patient names or other patient details with corresponding vector representations. Thus, simply distributing models across different institutes does not satisfy the privacy-preserving condition since it contains PHI information.[1]

Recently, data encryption techniques have been used to protect DL distribution and provide a secure collaboration environment. Techniques, such as differential privacy [20] and homomorphic encryption [21], have shown the ability to protect model shared parameters from re-identification attacks. The challenge of encryption techniques is that they can be attacked by untrusted platforms and unauthorized users. Some of them are not robust and can be affected by noise and errors. Moreover, the protection mechanism has to consider computation resource requirements, such as computation time, memory usage, etc. Encryption-based tools require additional computational cost, which may raise resource costs above acceptable levels. Most existing research on differential privacy in DL focuses either on preventing users from gaining knowledge about the training data when the model is deployed in the inference stage [20], [22], [23] or on how to train a model on multiple datasets without directly sharing access to those datasets [20], [24]. However, differential privacy often comes at the cost of an accuracy reduction for models trained on the corrupted data [20]. Also, information can be retrieved from a trained model using adversarial networks even when differential privacy mechanisms have been applied [25]. Khattak et al. [26] have presented a survey of word embeddings for clinical text, and discussed the limitations of word embeddings including privacy issues for clinical data.

This work introduces a simple and inexpensive method to prevent PHI information from entering the word embedding vocabulary. It offers a privacy by design solution without losing in accuracy performance nor increasing the computation cost. By combining this method with TL techniques such as cyclic or acyclic training, we show that users can gain the performance benefits of training on additional sensitive data without directly accessing that data. We demonstrate the effectiveness of our approach in the clinical application of classifying key data elements in cancer pathology reports in which protected data is spread across multiple cancer registries.

1. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

3 Materials and Methods

3.1 Datasets and Pre-processing

We used text corpora of cancer pathology reports obtained from the Louisiana Tumor Registry (LTR) and Kentucky Cancer Registry (KCR) of the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) Program. The study was executed in accordance to the institutional review board protocol DOE000152. The LTR and KCR datasets consist of 374,899 and 172,128 pathology reports respectively. The LTR corpus spans the period 2004-2018 while the KCR corpus spans the period 2009-2018. Each pathology report is identified by a combination of patient ID and tumor ID, which is called case ID. Each case ID may be associated with one or more pathology reports. Certified Tumor Registrars (CTRs) 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[2]. Labels were provided for various data elements, such as tumor type, and other cancer characteristics. In this paper, we consider the International Classification of Diseases for Oncology, Third Edition (ICD-O-3), topography (i.e., site/subsite), laterality, behavior, histology, and grade as the data elements of interest as they are fundamental information extraction tasks for cancer reporting. Figures 8 and 9 in Appendix A show the number of occurrences per label of all six cancer characteristics in LTR and KCR datasets, respectively. We can see from the figures there is an extreme class imbalance. Some classes are represented by less than 10 pathology reports, while others are represented by thousands of pathology reports. Documents generated within 10 days between the date of diagnosis and either path specimen collection date or the surgery date were identified as relevant to the specific case ID. The 10-day window was based on an analysis of the pathology report submissions with the vast majority of reports and addenda included within that time frame. The remaining pathology reports which were outside the 10-day window were excluded from the study.

To simulate a scenario in which four cancer registries are interested in collaborating, we randomly split each registry corpus into two subsets with similar size which resulted into four separate datasets. Table 1 summarizes the total number of pathology reports for each "virtual" cancer registry and the number of labels observed in the corresponding data. For each set, 80% of the samples were used for train and validation with 80-20 ratio, while the remaining 20% was used for testing. Since multiple cancer pathology reports might have the same case ID, we ensured that unique case IDs can be either in train, validation or test sets to avoid any positive bias in the reported results.

We applied standard text pre-processing techniques to clean our corpus, as we have described in previous studies [4], [5], [7]. After excluding metadata (e.g., patient ID, registry ID) in cancer pathology reports, text was cleaned by removing any consecutive punctuation and by lowercasing all alphabetical characters. To reduce the vocabulary size, all words with document frequency less than five were replaced with an "$unknown\_word$" token, all decimals were

2. https://seer.cancer.gov/tools/codingmanuals/index.html

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

converted to a "$decimal$" word token, and all integers larger than 100 were converted to a "$large\_integer$" word token. Cancer pathology reports are represented as one dimensional vectors, where each element is a word token. Different lengths of cancer pathology reports are accommodated by specifying a fixed length of $L = 1,500$ words for all reports. All documents longer than L are truncated and all documents shorter than L are padded. Please note that 95% of the pathology documents in our dataset have fewer than 1,500 words.

3.2 Multitask CNN for Information Extraction from Text Data

Multitask learning (MTL) is a mechanism for learning multiple related tasks simultaneously while leveraging knowledge across the tasks [27]. These related tasks can be learned using the same or different datasets. MTL was successfully used to train a word-level convolutional neural network (CNN) model to extract simultaneously five different data elements from cancer pathology reports – site, laterality, behavior, histology, and grade [7]. In this approach, each data element of interest is modeled as a separate learning task. The common architecture of MTL utilizes shared hidden layers for all tasks and then one separate output layer for each task. For NLP applications, the first hidden layer of a CNN model is the embedding layer which represents the semantic meanings of words using d-sized real-valued vectors.

The word embeddings layer produces a 2-D document matrix of size ($L \times d$), where L is the document length. This matrix serves as input to the convolution layers. For NLP applications, the convolution layers in CNNs are not stacked as in computer vision models. Instead, they are structured as parallel layers that operate simultaneously on document matrices. Convolution filters are applied to the document matrix by sliding linear filters over the text in order to extract features at each position. To extract multiple features, multiple filters are used with variable window sizes. Since words are represented by d-sized vectors, the width of filters equals to d. Thus, the size of filters is $n \times d$, where the height of a filter n corresponds to a context length of n word vectors or an n-gram. Convolution layers with non-linear activations generate L-sized feature maps which are the representation of every context window over the document matrix. Then, a max pooling layer is added to capture the most important features by taking the max value from each feature map as the extracted feature from a particular filter. The outputs of the pooling layers are concatenated by the last layers shared across all tasks. These shared layers are followed by multiple, fully connected, task-specific softmax layers to produce a rank for each label. Each task has a separate fully connected layer and its size is determined by the number of labels for each task.

In this paper, we adopted MTL to train a CNN model to extract six different cancer characteristics from cancer pathology reports: site, subsite laterality, behavior, histology, and grade. Figure 1 illustrates the architecture diagram of the MT-CNN model used in this paper. The network weights are trained using the ADADELTA adaptive gradient descent algorithm treating the loss weight for all tasks equally as in [28]. Dropout was applied with probability 50%. The number of filters in each set was 300, and the kernel sizes, K1, K2, and K3 are 3, 4, and 5, respectively. These parameters were optimized following our previous studies [4], [7]. The model hyper-parameters were initialized as the architecture presented in [29]. Then, we specified the search space of the substantial hyper-parameters to be explored. We used $Scikit\text{-}Optimize$ library methods to find the best hyper-parameters.

3.3 Word Embeddings

Word embeddings have been recognized as one of the key breakthroughs for various NLP applications such as document classification [30], and machine translation [31]. Word embeddings provide a way of converting words into numerical vectors which are used as inputs to DL models. These vectors have relatively lower dimensional features than the one-hot representation. Word embeddings have been shown to capture semantic information via observed similarities in word contexts, where the vector representations of semantically similar words are close to each other. Thus, they insert contextual knowledge into models helping DL algorithms to automatically understand word analogies and capture their semantic properties [32].

Figure 2 illustrates the traditional word embeddings process. It starts by collecting all unique words in a corpus as a vocabulary list of size V. Then each word in the vocabulary list is assigned to an integer index i, where $i \in \{1, 2, \dots V\}$. The vocabulary is saved in a dictionary format, where keys are the word tokens and values are their indices. For each document of size L in the dataset, the words are converted to their corresponding indices using the vocabulary dictionary. These indices are used to access the corresponding word vector representations in the embedding LUT. The number of embedding LUT parameters is proportional to the vocabulary size and word vector representation length, i.e. if a text corpus has V unique words and the feature representation of each word is a d sized vector, then the embedding LUT is going to be $d \times V$ dimensional, and each word has a notation that corresponds to d by a one-dimensional embedding vector. This process results in a document matrix of size $L \times d$ which is used as input to the convolution layer. Since the dictionary is associated to the NLP DL model trained on the data corpus, it is required when the trained model is used for inference. Since this dictionary is comprised of word tokens that appear in the data corpus of the cancer registry that provides the training data, it is expected to include word tokens associated with patient last names and other protected identifier information. Thus, the trained model and its related dictionary does not preserve data privacy if it is shared with another cancer registry.

The word vector representations can be learned from a large text corpus through Word2Vec [33] or GloVe [34] techniques separately from the other model parameters. They can also be learned from a task specific dataset with the other model parameters through back propagation. In this paper, the word embeddings parameters are randomly initialized and learned through back propagation since previous studies have shown to work well for this application [4].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

TABLE 1
Statistics of LTR, KCR and the centralized datasets and label counts for each cancer key characteristic.

Dataset	Train	Validation	Test	Total	site	subsite	laterality	behavior	histology	grade
LTR dataset-1	83,172	20,858	26,007	130,037	70	295	7	4	464	9
LTR dataset-2	83,418	20,769	26,019	130,206	70	294	7	4	463	9
KCR dataset-1	47,862	11,984	14,884	74,730	69	290	7	4	428	8
KCR dataset-2	48,109	12,015	14,967	75,091	69	287	7	4	428	8
Centralized	262,561	65,626	81,877	410,064	70	313	7	4	543	9

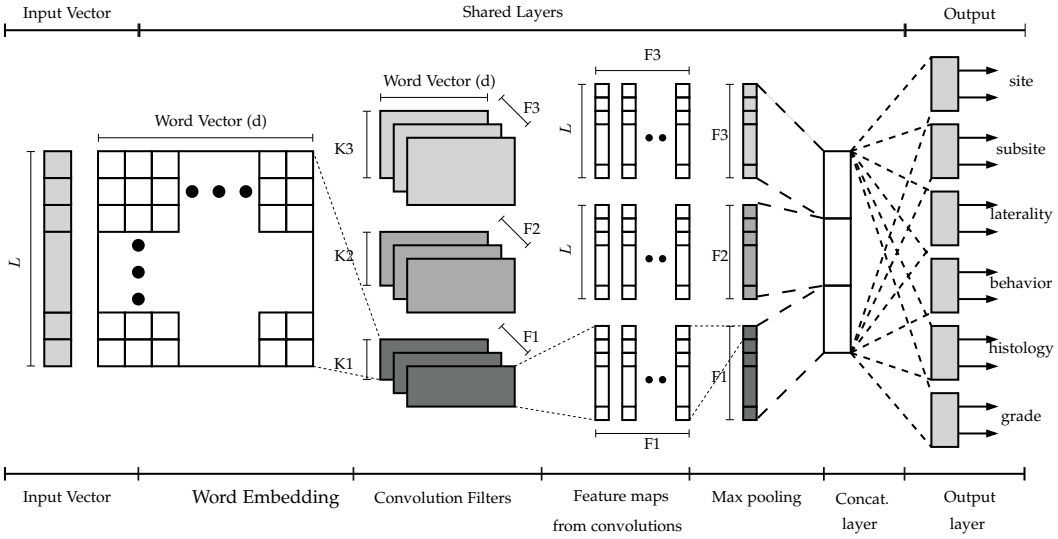

Fig. 1. Architecture diagram of the MT-CNN, where (F1, F2, F3) are the number of filters in each convolution layer, while (K1, K2, K3) are the kernel size for each set of filters. In this paper, (F1, F2, F3) are 300 filters each, (K1, K2, K3) are (3, 4, 5) respectively, Word Vector (d) is 300, and L is 1500 word tokens.

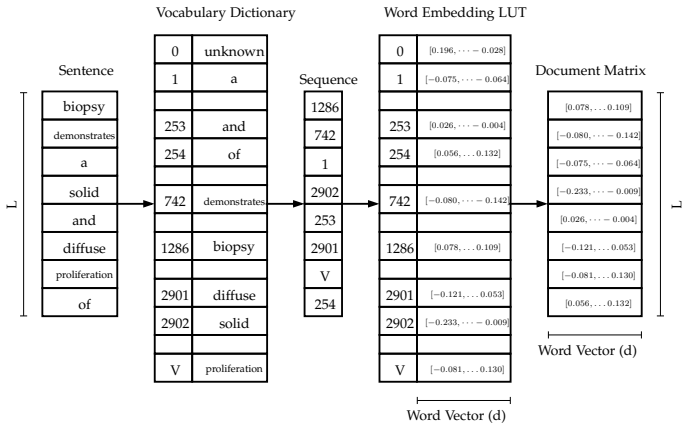

Fig. 2. Word embedding example diagram, where vocabulary dictionary converts words in the input sentence to the corresponding indices and V is the vocabulary dictionary size.

For the models without privacy-preserving, the dictionary created using the corpus of one cancer registry is shared with other registries. However, for the privacy-preserving models, only word tokens that appear in publicly available word embeddings are shared across registries. Although there are many available embeddings that are trained on public datasets, we have used the embeddings trained on PubMed and MIMIC-III datasets [35]. Unlike public datasets such as Wikipedia and Google news, which may include patient names, residential addresses, etc. that match what registries have in their private data, the vocabulary of MIMIC-III and Pubmed do not contain PHI information. We note that the vocabulary of MIMIC-III and Pubmed covers about 83% of the word tokens appearing in the LTR and KCR datasets.

3.4 Transfer Learning

Transfer learning is defined as the process of transferring knowledge learned from a source task, which can be a dataset, to a target task [36]. It can be done either by transferring the low-level layers [37], the high-level layers [38], or the whole model layers [9]. More details about a complete study on the impact of layer transferability can be found in [39]. The need for transfer learning emerges when the labeled training dataset for a model cannot be shared due to data sharing restrictions or when the dataset available for learning is limited or highly imbalanced. This is the case with population cancer registries in which not only there are privacy concerns regarding patient data sharing but also cancer registry data demonstrate substantial imbalance as some cancer types are extremely rare while other cancer types are very common. Transfer learning has been successfully applied to different computer vision applications, such as image classification [8] including clinical imaging applications [40], [41]. In such applications, computer vision

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

models pre-trained on a very large but general image data (e.g., ImageNet) are exploited to transfer knowledge to a specialized clinical imaging dataset which is relatively small but sufficient for domain-driven fine-tuning of the general trained model. The success of applying transfer learning on image applications, opened up the possibility to exploit transfer learning in non-clinical NLP applications, such as sentiment classification [42]. However, applying transfer learning of DL models to clinical NLP tasks is still an understudied research topic.

For NLP applications, word embeddings pre-trained on unlabeled corpora using unsupervised learning approaches, such as Word2Vec or GloVe have been extensively used to transfer knowledge across tasks. This approach was also used successfully for clinical NLP tasks by transferring the embeddings of medical concepts learned from multimodal medical data [43]. However, this approach did not improve the performance of CNN models for information extraction from cancer pathology reports [4]. This finding suggests that transferring knowledge from general datasets (e.g., Google News, PubMed) that are not semantically similar to the target dataset (pathology reports) does not produce the best performing models.

The main obstacle of applying transfer learning across cancer registries to tackle text information extraction tasks is preserving privacy. Transferring a model trained on a cancer registry corpus requires transferring its associated word embeddings and dictionary, which holds patient identifiers. Therefore, sharing trained NLP models across cancer registries presents unique challenges. To evaluate the extent of the problem, we implement three transfer learning approaches using the MT-CNN as the base NLP model: 1) The conventional transfer learning without privacy-preserving by transferring the whole model parameters across registries; 2) Transfer learning with privacy-preserving by dropping the embedding layer parameters and sharing the remaining model parameters; and 3) A novel privacy-preserving transfer learning approach by constructing vocabulary dictionary from word tokens available in publicly accessible pre-trained word embeddings (instead of using all word tokens appearing in the cancer registry corpus).

4 Experiments

4.1 Experimental Setup

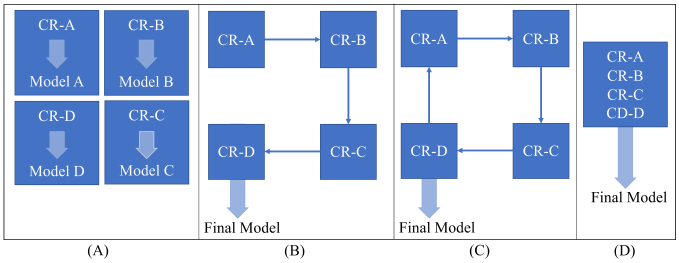

Fig. 3. DL model training configurations: A) Single-registry model, B) acyclic transfer learning with/without privacy-preserving, C) cyclic transfer learning with/without privacy-preserving, D) centralized model. Where CR is Cancer Registry.

In this paper, we perform comparative analysis of various transfer learning MT-CNN models for extracting six key cancer characteristics from pathology reports – site, subsite, laterality, behavior, histology, and grade. Specifically, we explored five different transfer learning approaches: (i) transfer learning with drop embeddings model, (ii) acyclic transfer learning without privacy preserving model, (iii) cyclic transfer learning without privacy preserving model, (iv) acyclic transfer learning with privacy-preserving model, and (v) cyclic transfer learning with privacy-preserving model. We benchmarked these models relative to single registry models and a centralized model. Below we describe the various models starting with the models that offer the highest privacy protection:

- **Single-registry model:** This is the baseline model. A MT-CNN model is trained and tested on each registry separately without sharing any information across them. This approach offers the highest data privacy and protection since nothing is shared across cancer registries. However, the limited dataset size available in each dataset may affect overall model accuracy.
- **Transfer learning with drop embeddings:** This approach was implemented to study the importance of sharing word embeddings relative to the other model parameters across cancer registries. A MT-CNN model is trained on one of the registry datasets. Then, the trained parameters, excluding the embeddings, are transferred to the next registry. This approach offers a privacy-preserving property since the vocabulary dictionary is not released to other registries.
- **Acyclic transfer learning with privacy preserving:** Acyclic distribution is the traditional transfer learning approach. A model is trained at one cancer registry and then it is shared with the next registry for further fine-tuning. The process can continue across all collaborating registries, each one fine-tuning the shared model with their own data. This approach opens questions whether the acyclic transfer learning model differs depending on the registry order for model sharing and fine-tuning. To ensure privacy preservation, we build a vocabulary dictionary from all available word tokens in the registry training corpus while excluding all word tokens that are not available in a publicly available vocabulary dictionary (as described in the previous section).
- **Cyclic transfer learning with privacy preserving:** Cyclic model distribution was proposed by K. Chang et al. [10] and used for medical imaging applications. In cyclic distribution, a MT-CNN model is trained for a certain number of epochs on one of the registries. Then, the model is transferred to the next registry for further training with local data. This process is iterated in a cyclic manner across all collaborating registries until the model converges. The resultant model is shared across registries for testing purposes. To ensure privacy-preserving, the vocabulary dictionary is built from the corpus word tokens that are available in a publicly available vocabulary dictionary as in the acyclic transfer learning with privacy preserving approach.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

- **Acyclic transfer learning without privacy preserving:** This approach is similar to acyclic transfer learning with privacy preserving in terms of sequentially distributing the model across cancer registries without iteration. However, this approach builds the vocabulary dictionary from all words observed in the training corpus without restrictions. Since the vocabulary dictionary is shared across registries, this approach offers less data privacy though it does not directly associate a first name with a specific last name or cancer type. Still, this approach presents increased risk for reverse engineering since Protected Health Information (PHI) information can be captured from these tokens.
- **Cyclic transfer learning without privacy preserving:** This approach is similar to cyclic transfer learning with privacy preserving in terms of model sharing across cancer registries in a cyclic manner. However, the vocabulary dictionary is built as in the acyclic transfer learning without privacy preserving. Thus, it offers less privacy for the same reasons described above.
- **Centralized model:** In this approach, pathology reports are collected from all collaborating cancer registries and hosted in a centralized location. Then, a global MT-CNN model is trained on the whole corpus and shared with cancer registries for testing. This approach offers the lowest data privacy and protection among other approaches due to data sharing with the central hub. It is expected though that this approach also offers the best classification accuracy as all the data is aggregated to train a global model.

4.2 Performance Evaluation

We evaluate the models using standard NLP metrics – micro- and macro-F1 scores – for each of our six classification tasks. The micro-averaged metric is equivalent to the model performance accuracy. It assigns weight to each class proportional to the class prevalence in the dataset. This metric is not sufficient to evaluate model performance, especially when the dataset has extreme class imbalance. Therefore, macro-averaged F1-score is used to help in evaluating model performance on the less prevalent classes. The macro-F1 score gives equal weight to each class without considering the class size. The performance evaluation metrics of each task are calculated separately. For each class (i) in C, where C is the total number of classes and i $\in \{1, \ldots, |C|\}$, the number of class true positives, false positives, and false negatives are denoted TP(i), FP(i), and FN(i), respectively. Class-based metrics are defined as:

$$\text{Precision(i)} = \frac{\text{TP(i)}}{\text{TP(i)} + \text{FP(i)}}$$

$$\text{Recall(i)} = \frac{\text{TP(i)}}{\text{TP(i)} + \text{FN(i)}} \quad (1)$$

$$\text{F1-score(i)} = \frac{2 \times \text{Precision(i)} \times \text{Recall(i)}}{\text{Precision(i)} + \text{Recall(i)}}$$

$$\text{Macro-F1 score (i)} = \frac{1}{|C|} \cdot \sum_{i \in C} \text{F1-score(i)} \quad (2)$$

For all metrics, we calculate 95% confidence intervals by bootstrapping [44] from the test set to estimate the variability of a model performance metric. The confidence intervals are used to determine the statistical significance of the difference in model performance. If the confidence interval of a proposed model's performance metric has no overlap with the confidence interval of a baseline model's performance metric, then the two models are considered statistically significantly different. Algorithm 1 shows how to derive confidence intervals using the bootstrap procedure.

Algorithm 1: Bootstrapping Procedure for Confidence Interval

Input: y_true, y_pred, performance metric
Output: 95% confidence interval of the metric

```
bootstrap_samples = [];
for i = 1 to N do
    R = random samples with replacement of size y;
    y_true_bootstrap = y_true[R];
    y_pred_bootstrap = y_pred[R];
    F1-score(y_true_bootstrap,y_pred_bootstrap) ;
     // micro or macro
    Append score to bootstrap_samples;
percentile(bootstrap_samples, 2.5);
percentile(bootstrap_samples, 97.5);
```

5 Results

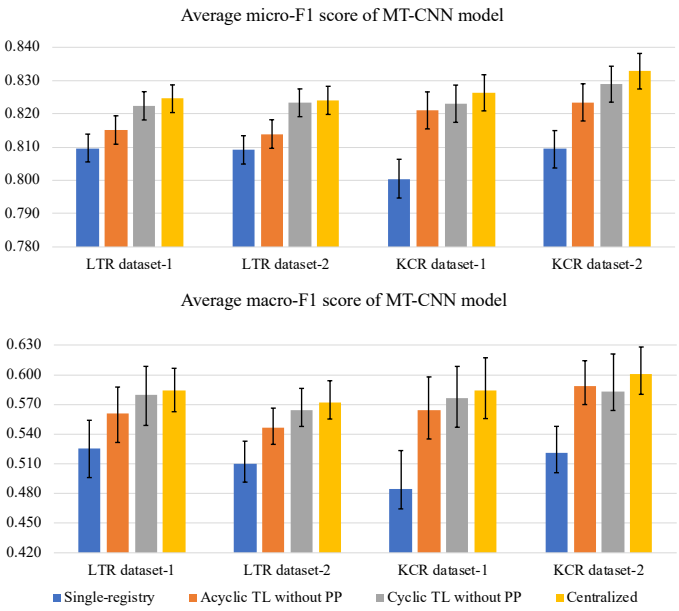

Fig. 4. Performance evaluation of different MT-CNN models without privacy-preserving training as compared to single-registry training (with 95% confidence intervals).

We use the average micro- and macro-F1 scores across all tasks to summarize the effectiveness different training approaches. Detailed experimental results are available in Tables 2 and 3 in Appendix B, which show the micro- and macro-F1 scores of the MT-CNN using different training approaches on each individual task – site, subsite, laterality, behavior, histology, and grade.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

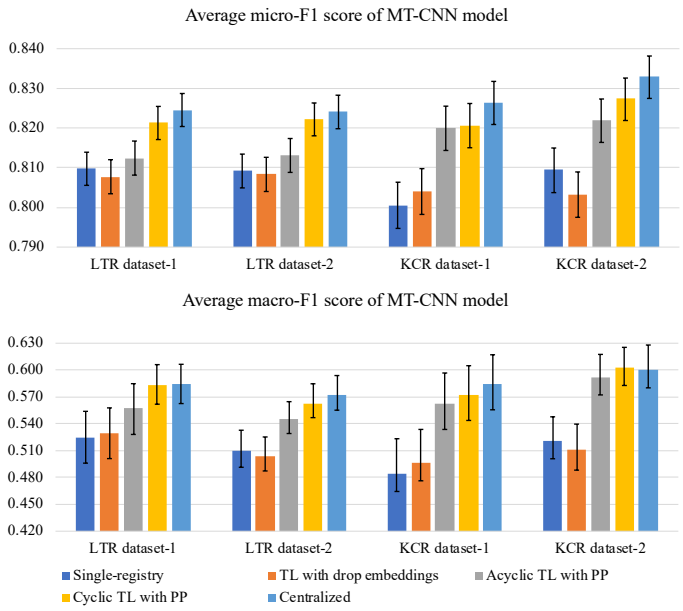

Fig. 5. Performance evaluation of different MT-CNN models with privacy-preserving training as compared to single-registry and centralized training methods (with 95% confidence intervals).

In our first set of experiments, we evaluate the effectiveness of collaboration methods among cancer registries without any privacy-preserving considerations – these methods include centralized learning, cyclic TL without privacy-preserving, and acyclic TL without privacy-preserving. We compare these training approaches with single-registry learning in terms of micro- and macro-F1 scores across all pathology report datasets – LTR dataset-1, LTR dataset-2, KCR dataset-1, and KCR dataset-2 – as shown in Figure 4. Across all datasets, the single-registry model has the lowest performance as compared to other approaches. Specifically, the average micro- and macro-F1 scores are: LTR dataset-1 (0.810, 0.525), LTR dataset-2 (0.809, 0.510), KCR dataset-1 (0.800, 0.484), and KCR dataset-2 (0.809, 0.521). The inferior performance of the single registry model highlights the importance of collaboration among cancer registries by leveraging each other's data. The centralized model, which is concurrently trained on data from all registries, achieves a statistically significantly better performance across all datasets compared to the single-registry model. The centralized model achieves average micro- and macro-F1 scores on LTR dataset-1 (0.825, 0.584), LTR dataset-2 (0.824, 0.572), KCR dataset-1 (0.826, 0.584), and KCR dataset-2 (0.833, 0.601). The performance improvement is particularly notable for the macro-F1 scores highlighting the performance gains for the low prevalence classes as the dataset size and variability in cancer reports increase from single-registry data to multiple-registry data. The centralized model performance serves as the ideal case baseline for the other models to reach while preserving patient privacy and without data sharing. Acyclic transfer learning without privacy-preserving significantly outperforms the baseline single-registry model with average micro- and macro-F1 scores of: LTR dataset-1 (0.815, 0.560), LTR dataset-2 (0.814, 0.546), KCR dataset-1 (0.821, 0.564), and KCR dataset-2 (0.823, 0.589). This approach performs well for many tasks and datasets compared to the centralized model with a marginal drop in performance which is not statistically significant. However, there is a significant degradation in performance for LTR dataset-2 micro-F1 score. Tables 2 and 3 in Appendix B show the drop in performance for some tasks, such as subsite in LTR dataset-1 and grade in KCR dataset-2. Cyclic transfer learning without privacy-preserving approach appears to mitigate this performance drop, outperforming the acyclic transfer learning without privacy-preserving approach across all datasets and information extraction tasks with average micro- and macro-F1 scores of: LTR dataset-1 (0.822, 0.580), LTR dataset-2 (0.823, 0.565), KCR dataset-1 (0.823, 0.576), and KCR dataset-2 (0.829, 0.583). This approach also reaches the performance level of the centralized model.

The second set of experiments compare the privacy-preserving approaches – acyclic and cyclic transfer learning with privacy-preserving and transfer learning with drop embeddings – with the centralized and single-registry training. Figure 5 shows the performance of these training approaches across all datasets. We also compare acyclic and cyclic TL with and without the privacy-preserving consideration. The most straightforward transfer learning with PHI privacy-preserving approach is to drop the word embeddings and share the remaining model parameters (i.e., transfer learning with drop embeddings model). The average micro- and macro-F1 scores of this approach are very close to the single-registry model performance: LTR dataset-1 (0.808, 0.530), LTR dataset-2 (0.808, 0.504), KCR dataset-1 (0.804, 0.496), and KCR dataset-2 (0.803, 0.512). This finding makes intuitive sense since the convolution parameters are associated with the embeddings and trained to capture features in the specific embeddings. Dropping the embeddings from the model may help preserve privacy but it does not transfer useful knowledge across cancer registries and does not provide any performance gains over the single-registry model, which is the main reason for collaboration among registries. The acyclic transfer learning with privacy-preserving model shows some benefits over the single-registry training for some datasets, but not the others. It achieves average micro- and macro-F1 scores on LTR dataset-1 (0.812, 0.557), LTR dataset-2 (0.813, 0.545), KCR dataset-1 (0.820, 0.563), and KCR dataset-2 (0.822, 0.592). On the other hand, the cyclic transfer learning with privacy-preserving model outperforms the single-registry model across all datasets. It achieves average micro- and macro-F1 scores on LTR dataset-1 (0.821, 0.584), LTR dataset-2 (0.822, 0.563), KCR dataset-1 (0.821, 0.572), and KCR dataset-2 (0.827, 0.603). Finally, both the acyclic and cyclic transfer learning with privacy-preserving models attain the same performance as the acyclic and cyclic transfer learning without privacy-preserving models, with micro- and macro-F1 scores falling within the confidence intervals across all tasks. Since the cyclic method is consistently better than the acyclic one across all tasks, it is deemed as the best choice. Compared to the centralized model, the cyclic transfer learning with privacy preserving is statistically indistinguishable and in some cases even superior.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

6 Discussion

Our experiments show that data and model sharing approaches among cancer registries consistently improve the performance of a MT-CNN NLP model for information extraction from cancer pathology reports as compared to the single-registry model. This finding highlights the importance of collaboration across cancer registries to develop a more efficient DL model for NLP tasks. Transfer learning approaches, with and without privacy-preserving significantly outperform the transfer learning with drop-embeddings model. This is mainly due to the importance of sharing the embedding layer parameters along with other model parameters.

Figure 6 shows the validation loss convergence of the transfer learning approaches with and without privacy preservation. Although the acyclic model distribution yielded an increase in the speed of convergence of the final model, the cyclic distribution yielded a lower final loss value. The higher accuracy of cyclic model distribution indicates the model is not overfitting on one cancer registry dataset. Instead, it is more generalizable through the frequent model distribution among cancer registries. The superior performance is consistent across all information extraction tasks. Moreover, the cyclic model distribution – with or without privacy preserving – achieves comparable performance to the centralized model. A clear advantage of cyclic over acyclic transfer learning is that the cyclic approach is agnostic to the sequence in which a model is trained on one registry before it is distributed to the next one for fine-tuning. Based on additional experiments using acyclic transfer learning with different registry sequences, we observed variable model performance. Our study presented the performance of best acyclic transfer learning model. The general trend was that the cyclic model distribution is superior to the various acyclic transfer learning models in terms of performance.

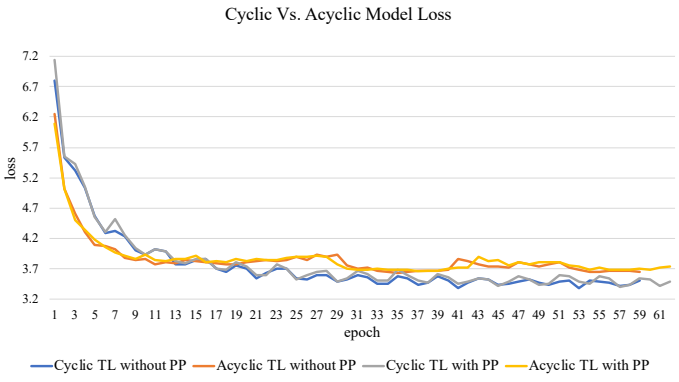

Fig. 6. Validation loss of transfer learning with and without privacy-preserving comparing cyclic and acyclic model distribution, where TL is transfer learning and PP is privacy-preserving.

The experiments demonstrate that our proposed transfer learning with privacy-preserving technique achieves comparable results to the conventional transfer learning without privacy-preserving and centralized models. For some tasks, although the performance metric is within the confidence interval, the difference is noticeable. Upon evaluation of the cases in which the difference is more than 1% we observed it is due to labels with very few samples. Model performance is not robust in low prevalence class labels, due to the extreme class imbalance in cancer registry data. For example, the ratio of low prevalent to high prevalent classes ranges from 1:64,898 to 1:6,489 for histology task in the LTR dataset. If we exclude the low prevalence class labels, the difference is reduced significantly. For example, the difference between the macro-F1 score of the cyclic transfer learning with privacy-preserving model and the centralized model on the histology task from LTR dataset-2 is 2.3%. When excluding the classes with fewer than 10 samples, this difference is reduced to 0.5%. The same trend is observed when comparing the cyclic transfer learning with privacy-preserving model to the cyclic transfer learning without privacy-preserving model on the laterality task for the KCR dataset-2. By excluding $laterality\_label = 3$, which has only one sample in the test set, the macro-F1 score difference reduces from 7.9% to 0.7%.

Besides the advantage of developing a better performing model by pulling the data from multiple cancer registries, data sharing and transfer learning can also help tackle the class imbalance problem. In cancer registries, this is a common challenge as some cancer types are highly prevalent (e.g., breast, lung, prostate) while others are very rare (e.g., esophagus, gum, sinuses). Figure 7 shows the performance of different training approaches on the histology task for prevalent histologies with at least 100 samples and for rare histologies with less than 100 samples. We selected the histology extraction task for illustration purposes since it has the highest number of class labels as well as the highest class-imbalance ratio. Please note though that the same trend is observed across all other information extraction tasks. As expected, all models perform relatively well on the more prevalent classes. However, on the low prevalence classes, the difference in performance is much clearer between the data sharing model (centralized), transfer learning models (acyclic transfer learning with and without privacy-preserving, cyclic transfer learning with and without privacy-preserving) and the non-transfer learning (single registry) or limited transfer learning models (drop-embeddings). Among the transfer learning approaches, cyclic learning with or without privacy-preserving achieves a performance comparable to the centralized model overcoming the challenge of imbalanced training data.

7 Conclusion

In this paper, we propose a privacy-preserving technique to share a DL NLP model across cancer registries, excluding any data that may compromise patient privacy. We demonstrate the value of our technique with a MT-CNN model for abstracting cancer characteristics from cancer pathology reports, a time-consuming manual activity across cancer registries. In addition, we study different model distribution and data sharing approaches with cancer registries. The experiments demonstrate that model distribution and data sharing approaches achieve the highest micro- and macro-F1 scores across all information extraction tasks, as compared to the single-registry model. The performance improvement is especially noticeable for macro-F1 scores,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

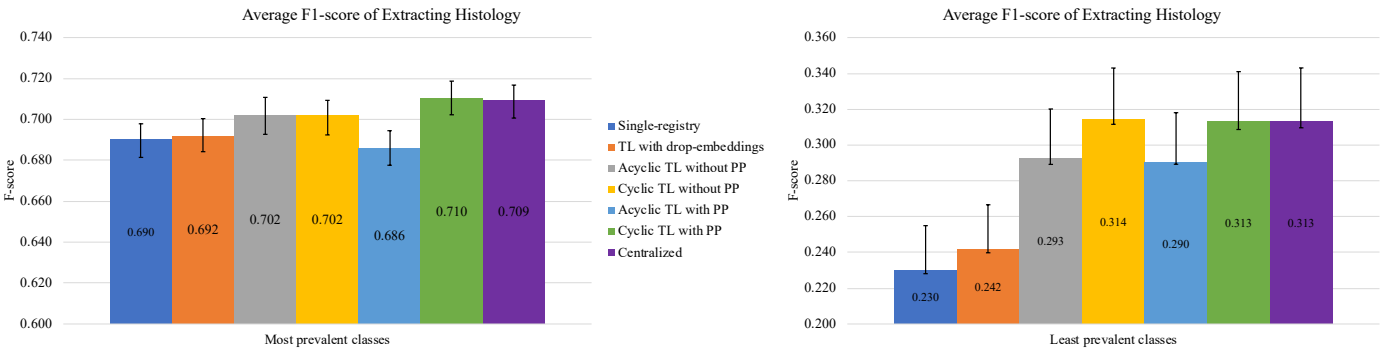

Fig. 7. Average F1-score of extracting the most prevalent classes (left) and least prevalent classes (right) histologies from cancer pathology reports using different training approaches, where TL is transfer learning and PP is privacy-preserving.

suggesting that these approaches do a better job classifying low prevalent cases which is an important advantage. Finally, our proposed transfer learning with privacy-preserving models achieve a comparable performance as the conventional transfer learning approach without privacy-preserving and the centralized model. This opens the possibility of sharing knowledge through NLP models across cancer registries without violating data privacy rules.

ACKNOWLEDGMENTS

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

This work has also been supported by National Cancer Institute under Contract No. HHSN261201800013I/HHSN26100001 and NCI Cancer Center Support Grant (P30CA177558).

REFERENCES

[1] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. F. Jones, R. Forshee, M. Walderhaug, and T. Botsis, "Natural language processing systems for capturing and standardizing unstructured clinical information," *J. of Biomedical Informatics*, vol. 73, pp. 14–29, Sept. 2017.

[2] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *Journal of Biomedical Informatics*, vol. 77, pp. 34–49, 2018.

[3] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.

[4] J. X. Qiu, H. J. Yoon, P. A. Fearn, and G. D. Tourassi, "Deep learning for automated extraction of primary sites from cancer pathology reports," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 244–251, Jan 2018.

[5] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan, "Hierarchical attention networks for information extraction from cancer pathology reports," *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 321–330, 2018.

[6] S. Gao, J. X. Qiu, M. Alawad, J. D. Hinkle, N. Schaefferkoetter, H.-J. Yoon, B. Christian, P. A. Fearn, L. Penberthy, X.-C. Wu, L. Coyle, G. Tourassi, and A. Ramanathan, "Classifying cancer pathology reports with hierarchical self-attention networks," *Artificial Intelligence in Medicine*, vol. 101, p. 101726, 2019.

[7] M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, and G. Tourassi, "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks," *Journal of the American Medical Informatics Association*, vol. 27, pp. 89–98, January 2020.

[8] A. Quattoni, M. Collins, and T. Darrell, "Transfer learning for image classification with sparse prototype representations," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[9] M. Alawad, S. Gao, J. Qiu, N. Schaefferkoetter, J. Hinkle, H. Yoon, J. Christian, X. Wu, E. Durbin, J. Jeong, I. Hands, D. Rust, and G. D. Tourassi, "Deep transfer learning across cancer registries for information extraction from pathology reports," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, May 2019.

[10] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, vol. 25, pp. 945–954, 03 2018.

[11] Y. Guo, R. Gaizauskas, I. Roberts, and G. Demetriou, "Identifying personal health information using support vector machines," in *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2006.

[12] F. Dernoncourt, J. Y. Lee, Ö. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *J Am Med Inform Assoc*, vol. 1;24(3), pp. 596–606, 2017.

[13] M. N. Sadat, M. M. A. Aziz, N. Mohammed, S. Pakhomov, H. Liu, and X. Jiang, "A privacy-preserving distributed filtering framework for NLP artifacts," *BMC Medical Informatics and Decision Making*, vol. 19, p. 183, Sep 2019.

[14] A. Stubbs, M. Filannino, and Z. Uzuner, "De-identification of psychiatric intake records," *J. of Biomedical Informatics*, vol. 75, pp. S4–S18, Nov. 2017.

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pp. 1273–1282, 2017.

[16] J. Chen, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous SGD," in *International Conference on Learning Representations Workshop Track*, 2016.

[17] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

Communications Security, CCS '17, (New York, NY, USA), pp. 603–618, ACM, 2017.

[18] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *CoRR*, vol. abs/1812.00564, 2018.

[19] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey, "No peek: A survey of private distributed deep learning," *CoRR*, vol. abs/1812.03288, 2018.

[20] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, (New York, NY, USA), pp. 308–318, ACM, 2016.

[21] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *Trans. Info. For. Sec.*, vol. 13, pp. 1333–1345, May 2018.

[22] C. Dwork and V. Feldman, "Privacy-preserving prediction," in *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pp. 1693–1702, 2018.

[23] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *CoRR*, vol. abs/1811.04017, 2018.

[24] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, (New York, NY, USA), pp. 1310–1321, ACM, 2015.

[25] Y. Li, T. Baldwin, and T. Cohn, "Towards robust and privacy-preserving text representations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Melbourne, Australia), pp. 25–30, Association for Computational Linguistics, July 2018.

[26] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics: X*, vol. 4, p. 100057, 2019.

[27] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017.

[28] M. Alawad, H. Yoon, and G. D. Tourassi, "Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pp. 218–221, March 2018.

[29] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014.

[30] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), pp. 649–657, MIT Press, 2015.

[31] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Association for Computational Linguistics, 2016.

[32] M. Alawad, S. M. S. Hasan, J. Blair Christian, and G. Tourassi, "Retrofitting word embeddings with the UMLS metathesaurus for clinical information extraction," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2838–2846, Dec 2018.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.

[34] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[35] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec: Improving Biomedical Word Embeddings with Subword Information and MeSH Ontology," *Scientific Data*, vol. 6, 05 2019.

[36] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, p. 9, May 2016.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, 2012.

[38] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3320–3328, 2014.

[40] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of digital imaging*, vol. 30, no. 2, pp. 234–243, 2016.

[41] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[42] T. Semwal, P. Yenigalla, G. Mathur, and S. B. Nair, "A practitioners' guide to transfer learning for text classification using convolutional neural networks," in *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA.*, pp. 513–521, 2018.

[43] A. L. Beam, B. Kompa, I. Fried, N. P. Palmer, X. Shi, T. Cai, and I. S. Kohane, "Clinical concept embeddings learned from massive sources of medical data," *CoRR*, vol. abs/1804.01486, 2018.

[44] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Taylor and Francis, 1994.

APPENDIX A
DATASET DISTRIBUTIONS

Figures 8 and 9 show the number of occurrences per label of all six cancer characteristics in LTR and KCR datasets, respectively.

APPENDIX B
RESULTS

Tables 2 and 3 illustrate the evaluation performance in terms of micro- and macro-F1 scores of different MT-CNN learning approaches on site, subsite, laterality, behavior, histology, and grade.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

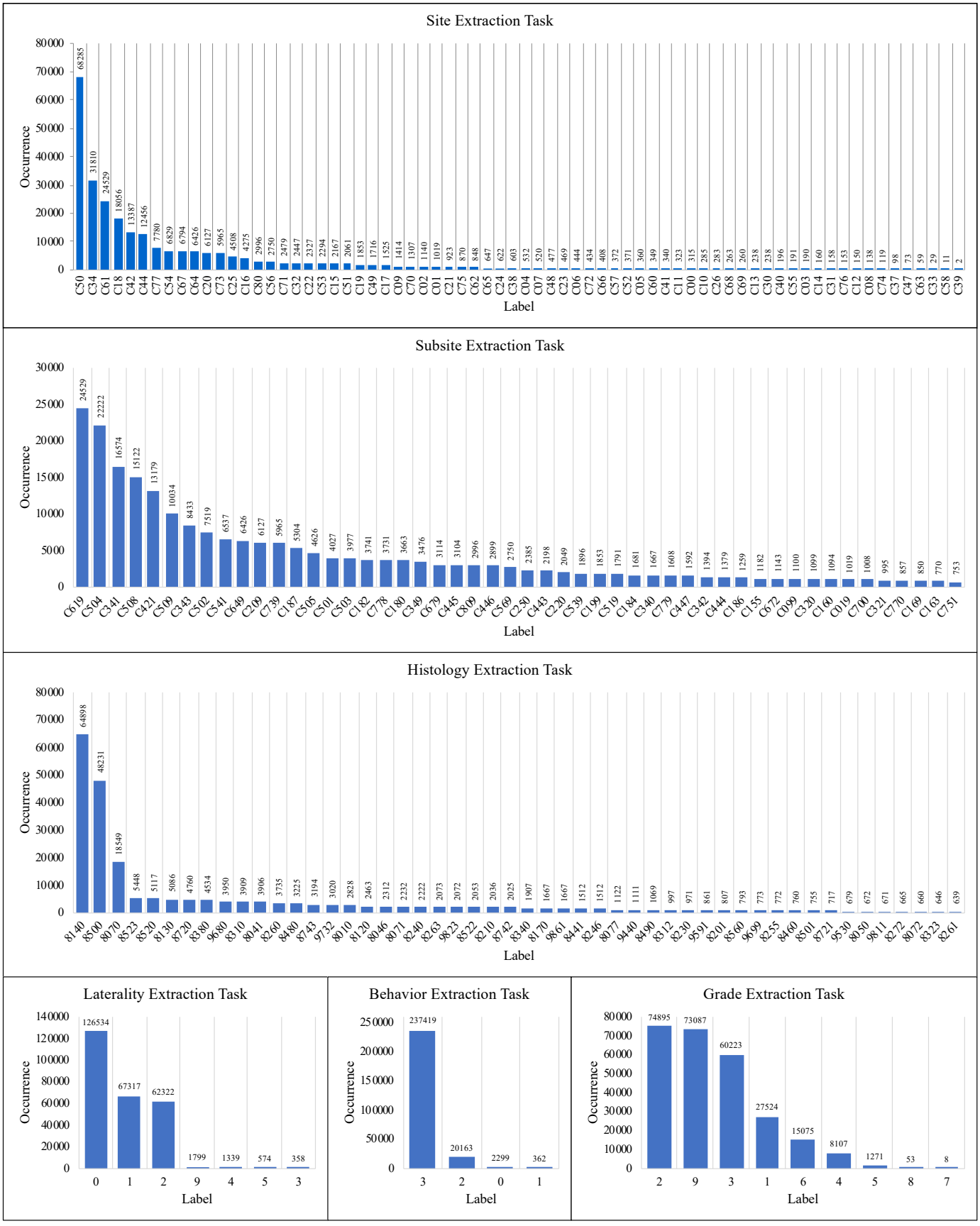

Fig. 8. Histograms of the number of occurrences per label of LTR dataset for each of the six classification tasks, arranged from most common to least common. For the subsite, and histology tasks, we only show the 50 most common labels. Detailed information about each label can be found online in the SEER coding manual at https:// seer.cancer.gov/tools/codingmanuals/.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

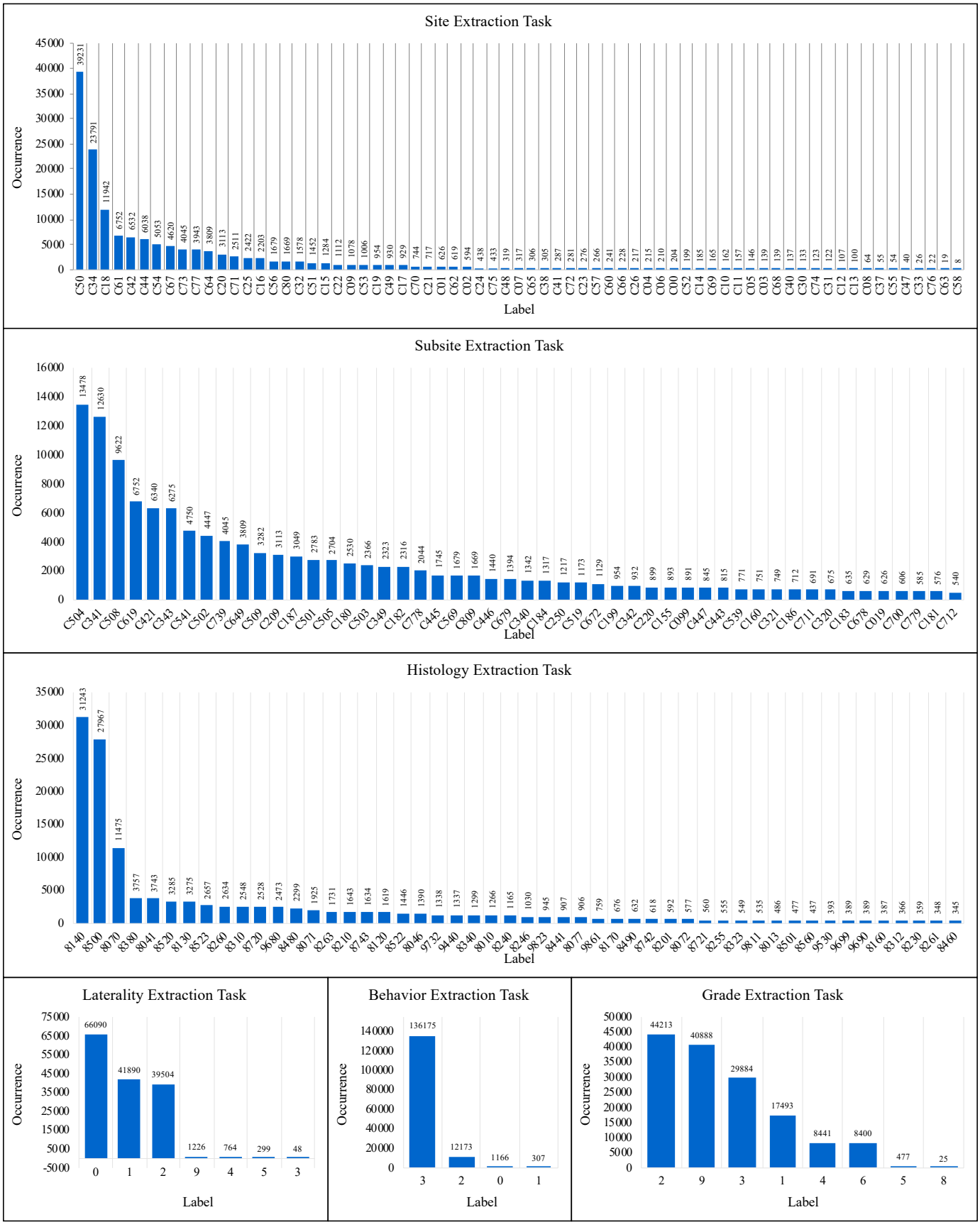

Fig. 9. Histograms of the number of occurrences per label of KCR dataset for each of the six classification tasks, arranged from most common to least common. For the subsite, and histology tasks, we only show the 50 most common labels. Detailed information about each label can be found online in the SEER coding manual at https:// seer.cancer.gov/tools/codingmanuals/.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

TABLE 2
Evaluation performance of different models on the site, subsite, and laterality extraction tasks (with 95% confidence intervals).

	LTR dataset-1		LTR dataset-2		KCR dataset-1		KCR dataset-2	
Model	Micro-F	Macro-F	Micro-F	Macro-F	Micro-F	Macro-F	Micro-F	Macro-F
site extraction task								
Single-registry	0.904 (0.901,0.908)	0.597 (0.584,0.621)	0.909 (0.905,0.912)	0.604 (0.583,0.619)	0.901 (0.896,0.906)	0.579 (0.560,0.608)	0.910 (0.906,0.914)	0.601 (0.581,0.626)
Transfer learning with drop-embeddings	0.904 (0.900,0.907)	0.591 (0.579,0.614)	0.906 (0.903,0.910)	0.594 (0.576,0.612)	0.906 (0.902,0.911)	0.618 (0.593,0.644)	0.905 (0.900,0.909)	0.602 (0.576,0.628)
Acyclic transfer learning without privacy-preserving	0.910 (0.906,0.913)	0.636 (0.621,0.658)	0.912 (0.909,0.916)	0.630 (0.610,0.649)	0.915 (0.911,0.920)	0.665 (0.640,0.685)	0.917 (0.913,0.921)	0.664 (0.638,0.689)
Cyclic transfer learning without privacy-preserving	0.914 (0.910,0.917)	0.669 (0.653,0.691)	0.917 (0.913,0.920)	0.661 (0.640,0.677)	0.918 (0.914,0.923)	0.685 (0.660,0.702)	0.919 (0.915,0.923)	0.679 (0.654,0.705)
Acyclic transfer learning with privacy-preserving	0.911 (0.907,0.915)	0.640 (0.622,0.663)	0.912 (0.908,0.915)	0.621 (0.600,0.640)	0.915 (0.910,0.919)	0.668 (0.644,0.688)	0.919 (0.914,0.923)	0.661 (0.636,0.687)
Cyclic transfer learning with privacy-preserving	0.914 (0.910,0.917)	0.637 (0.622,0.659)	0.918 (0.915,0.921)	0.655 (0.638,0.670)	0.918 (0.913,0.922)	0.674 (0.650,0.697)	0.919 (0.914,0.923)	0.675 (0.647,0.693)
Centralized	0.915 (0.912,0.918)	0.664 (0.647,0.687)	0.916 (0.912,0.919)	0.666 (0.644,0.680)	0.919 (0.915,0.923)	0.698 (0.672,0.717)	0.923 (0.918,0.927)	0.671 (0.650,0.702)
subsite extraction task								
Single-registry	0.605 (0.600,0.611)	0.287 (0.281,0.304)	0.599 (0.593,0.605)	0.279 (0.273,0.298)	0.585 (0.578,0.593)	0.264 (0.259,0.288)	0.598 (0.589,0.606)	0.264 (0.261,0.287)
Transfer learning with drop-embeddings	0.600 (0.594,0.606)	0.283 (0.276,0.299)	0.600 (0.594,0.606)	0.300 (0.290,0.317)	0.584 (0.576,0.592)	0.276 (0.272,0.301)	0.590 (0.582,0.598)	0.280 (0.273,0.304)
Acyclic transfer learning without privacy-preserving	0.623 (0.617,0.629)	0.326 (0.317,0.342)	0.621 (0.615,0.627)	0.323 (0.314,0.340)	0.635 (0.627,0.642)	0.343 (0.336,0.369)	0.626 (0.618,0.634)	0.343 (0.338,0.370)
Cyclic transfer learning without privacy-preserving	0.636 (0.630,0.642)	0.344 (0.334,0.360)	0.633 (0.627,0.639)	0.340 (0.332,0.360)	0.636 (0.629,0.644)	0.350 (0.345,0.379)	0.644 (0.637,0.651)	0.375 (0.364,0.400)
Acyclic transfer learning with privacy-preserving	0.614 (0.608,0.620)	0.324 (0.620,0.339)	0.616 (0.610,0.622)	0.318 (0.310,0.336)	0.621 (0.613,0.629)	0.331 (0.325,0.357)	0.619 (0.611,0.626)	0.350 (0.342,0.375)
Cyclic transfer learning with privacy-preserving	0.631 (0.625,0.637)	0.338 (0.329,0.355)	0.628 (0.622,0.634)	0.345 (0.338,0.367)	0.626 (0.618,0.634)	0.356 (0.349,0.381)	0.635 (0.627,0.642)	0.377 (0.365,0.400)
Centralized	0.640 (0.634,0.646)	0.362 (0.351,0.380)	0.634 (0.629,0.640)	0.359 (0.350,0.380)	0.638 (0.631,0.646)	0.357 (0.350,0.384)	0.643 (0.635,0.650)	0.371 (0.361,0.397)
laterality extraction task								
Single-registry	0.902 (0.899,0.906)	0.484 (0.467,0.500)	0.898 (0.894,0.901)	0.500 (0.470,0.531)	0.890 (0.885,0.895)	0.461 (0.432,0.488)	0.904 (0.899,0.908)	0.611 (0.581,0.639)
Transfer learning with drop-embeddings	0.900 (0.896,0.903)	0.499 (0.479,0.517)	0.900 (0.897,0.904)	0.463 (0.442,0.484)	0.896 (0.891,0.901)	0.445 (0.417,0.473)	0.904 (0.899,0.908)	0.607 (0.576,0.633)
Acyclic transfer learning without privacy-preserving	0.901 (0.897,0.905)	0.498 (0.480,0.515)	0.901 (0.897,0.905)	0.495 (0.473,0.514)	0.898 (0.893,0.903)	0.500 (0.472,0.525)	0.909 (0.905,0.914)	0.651 (0.622,0.677)
Cyclic transfer learning without privacy-preserving	0.907 (0.904,0.910)	0.525 (0.497,0.553)	0.908 (0.904,0.911)	0.515 (0.489,0.543)	0.902 (0.897,0.907)	0.486 (0.461,0.508)	0.914 (0.910,0.919)	0.539 (0.515,0.640)
Acyclic transfer learning with privacy-preserving	0.899 (0.896,0.903)	0.480 (0.461,0.499)	0.903 (0.899,0.906)	0.488 (0.467,0.508)	0.902 (0.897,0.907)	0.497 (0.469,0.523)	0.908 (0.903,0.913)	0.619 (0.590,0.647)
Cyclic transfer learning with privacy-preserving	0.904 (0.901,0.908)	0.534 (0.506,0.563)	0.903 (0.900,0.907)	0.505 (0.480,0.534)	0.899 (0.894,0.904)	0.488 (0.463,0.511)	0.910 (0.906,0.915)	0.618 (0.590,0.643)
Centralized	0.906 (0.902,0.909)	0.508 (0.491,0.525)	0.907 (0.903,0.911)	0.519 (0.494,0.549)	0.900 (0.895,0.905)	0.504 (0.479,0.527)	0.914 (0.910,0.919)	0.633 (0.606,0.658)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2983404, IEEE Transactions on Emerging Topics in Computing

TABLE 3
Evaluation performance of different models on the behavior, histology, and grade extraction tasks (with 95% confidence intervals).

	LTR dataset-1		LTR dataset-2		KCR dataset-1		KCR dataset-2	
Model	Micro-F	Macro-F	Micro-F	Macro-F	Micro-F	Macro-F	Micro-F	Macro-F
behavior extraction task								
Single-registry	0.973 (0.971,0.975)	0.798 (0.750,0.842)	0.974 (0.973,0.976)	0.781 (0.738,0.818)	0.963 (0.959,0.966)	0.790 (0.741,0.829)	0.968 (0.965,0.971)	0.791 (0.745,0.831)
Transfer learning with drop-embeddings	0.973 (0.971,0.975)	0.819 (0.775,0.859)	0.971 (0.969,0.973)	0.763 (0.724,0.800)	0.966 (0.963,0.969)	0.835 (0.792,0.868)	0.960 (0.957,0.963)	0.721 (0.673,0.766)
Acyclic transfer learning without privacy-preserving	0.974 (0.972,0.976)	0.857 (0.817,0.891)	0.974 (0.973,0.976)	0.857 (0.822,0.887)	0.970 (0.968,0.973)	0.839 (0.800,0.874)	0.970 (0.967,0.972)	0.872 (0.833,0.902)
Cyclic transfer learning without privacy-preserving	0.976 (0.974,0.978)	0.864 (0.823,0.896)	0.976 (0.974,0.978)	0.861 (0.829,0.890)	0.970 (0.968,0.973)	0.851 (0.815,0.881)	0.970 (0.968,0.973)	0.864 (0.826,0.896)
Acyclic transfer learning with privacy-preserving	0.976 (0.974,0.977)	0.863 (0.824,0.895)	0.976 (0.974,0.978)	0.877 (0.846,0.902)	0.972 (0.969,0.975)	0.846 (0.806,0.878)	0.972 (0.969,0.975)	0.892 (0.857,0.919)
Cyclic transfer learning with privacy-preserving	0.977 (0.976,0.979)	0.873 (0.833,0.905)	0.979 (0.977,0.981)	0.873 (0.841,0.902)	0.975 (0.972,0.977)	0.873 (0.837,0.902)	0.976 (0.973,0.978)	0.907 (0.875,0.933)
Centralized	0.979 (0.977,0.980)	0.852 (0.805,0.888)	0.978 (0.976,0.980)	0.877 (0.844,0.907)	0.974 (0.971,0.976)	0.881 (0.849,0.911)	0.973 (0.970,0.975)	0.865 (0.820,0.900)
histology extraction task								
Single-registry	0.748 (0.743,0.753)	0.266 (0.268,0.292)	0.751 (0.745,0.756)	0.278 (0.276,0.302)	0.729 (0.722,0.736)	0.216 (0.218,0.245)	0.730 (0.723,0.737)	0.230 (0.226,0.254)
Transfer learning with drop-embeddings	0.752 (0.746,0.757)	0.275 (0.276,0.301)	0.751 (0.745,0.756)	0.288 (0.286,0.312)	0.731 (0.724,0.738)	0.226 (0.227,0.253)	0.729 (0.722,0.736)	0.259 (0.252,0.283)
Acyclic transfer learning without privacy-preserving	0.762 (0.757,0.767)	0.323 (0.322,0.351)	0.757 (0.752,0.763)	0.345 (0.342,0.372)	0.752 (0.745,0.759)	0.321 (0.321,0.352)	0.761 (0.754,0.768)	0.347 (0.346,0.380)
Cyclic transfer learning without privacy-preserving	0.767 (0.762,0.772)	0.345 (0.344,0.374)	0.768 (0.762,0.773)	0.371 (0.366,0.398)	0.756 (0.749,0.763)	0.377 (0.367,0.406)	0.764 (0.757,0.771)	0.377 (0.369,0.405)
Acyclic transfer learning with privacy-preserving	0.758 (0.753,0.763)	0.322 (0.323,0.351)	0.755 (0.750,0.761)	0.342 (0.337,0.367)	0.751 (0.744,0.758)	0.325 (0.322,0.358)	0.761 (0.754,0.767)	0.377 (0.371,0.407)
Cyclic transfer learning with privacy-preserving	0.769 (0.764,0.774)	0.347 (0.346,0.375)	0.769 (0.764,0.774)	0.354 (0.349,0.383)	0.755 (0.748,0.762)	0.344 (0.340,0.375)	0.763 (0.757,0.770)	0.372 (0.365,0.403)
Centralized	0.771 (0.766,0.776)	0.347 (0.346,0.378)	0.771 (0.765,0.776)	0.377 (0.371,0.402)	0.765 (0.759,0.772)	0.359 (0.352,0.388)	0.774 (0.767,0.780)	0.395 (0.387,0.426)
grade extraction task								
Single-registry	0.726 (0.721,0.731)	0.717 (0.627,0.765)	0.725 (0.719,0.730)	0.618 (0.608,0.628)	0.735 (0.728,0.742)	0.595 (0.575,0.682)	0.747 (0.740,0.754)	0.630 (0.612,0.650)
Transfer learning with drop-embeddings	0.719 (0.713,0.724)	0.710 (0.622,0.758)	0.722 (0.716,0.727)	0.616 (0.605,0.626)	0.739 (0.733,0.746)	0.579 (0.556,0.663)	0.732 (0.725,0.739)	0.601 (0.580,0.623)
Acyclic transfer learning without privacy-preserving	0.721 (0.715,0.727)	0.722 (0.633,0.768)	0.718 (0.713,0.724)	0.626 (0.616,0.635)	0.757 (0.750,0.764)	0.714 (0.640,0.781)	0.758 (0.750,0.764)	0.655 (0.641,0.667)
Cyclic transfer learning without privacy-preserving	0.735 (0.729,0.740)	0.729 (0.641,0.777)	0.739 (0.734,0.745)	0.640 (0.630,0.649)	0.755 (0.748,0.762)	0.707 (0.633,0.774)	0.762 (0.755,0.769)	0.666 (0.653,0.679)
Acyclic transfer learning with privacy-preserving	0.716 (0.711,0.722)	0.714 (0.625,0.761)	0.717 (0.711,0.722)	0.624 (0.615,0.634)	0.760 (0.753,0.766)	0.710 (0.635,0.775)	0.753 (0.746,0.760)	0.652 (0.638,0.670)
Cyclic transfer learning with privacy-preserving	0.732 (0.727,0.738)	0.772 (0.735,0.779)	0.736 (0.731,0.741)	0.644 (0.635,0.653)	0.751 (0.745,0.758)	0.697 (0.622,0.763)	0.761 (0.754,0.767)	0.667 (0.654,0.680)
Centralized	0.737 (0.732,0.742)	0.772 (0.734,0.780)	0.738 (0.733,0.744)	0.637 (0.627,0.646)	0.762 (0.755,0.769)	0.707 (0.632,0.775)	0.771 (0.765,0.777)	0.670 (0.657,0.685)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.