Cancer Biomarkers 33 (2022) 185–198
DOI 10.3233/CBM-210306
IOS Press

Optimal vocabulary selection approaches for privacy-preserving deep NLP model training for information extraction and cancer epidemiology

Hong-Jun Yoon[a,*], Christopher Stanley[a], J. Blair Christian[a], Hilda B. Klasky[a], Andrew E. Blanchard[a], Eric B. Durbin[b], Xiao-Cheng Wu[c], Antoinette Stroup[d], Jennifer Doherty[e], Stephen M. Schwartz[f], Charles Wiggins[g], Mark Damesyn[h], Linda Coyle[i] and Georgia D. Tourassi[j]
[a]*Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[b]*College of Medicine, University of Kentucky, Lexington, KY, USA*
[c]*Louisiana State University Health Sciences Center, School of Public Health, New Orleans, LA, USA*
[d]*Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA*
[e]*Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA*
[f]*Fred Hutchinson Cancer Research Center, Epidemiology Program, Seattle, WA, USA*
[g]*University of New Mexico, Albuquerque, NM, USA*
[h]*California Department of Public Health, Sacramento, CA, USA*
[i]*Information Management Services Inc., Calverton, MD, USA*
[j]*National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA*

Received 1 June 2021
Accepted 11 September 2021

Abstract.
BACKGROUND: With the use of artificial intelligence and machine learning techniques for biomedical informatics, security and privacy concerns over the data and subject identities have also become an important issue and essential research topic. Without intentional safeguards, machine learning models may find patterns and features to improve task performance that are associated with private personal information.
OBJECTIVE: The privacy vulnerability of deep learning models for information extraction from medical textural contents needs to be quantified since the models are exposed to private health information and personally identifiable information. The objective of the study is to quantify the privacy vulnerability of the deep learning models for natural language processing and explore a proper way of securing patients' information to mitigate confidentiality breaches.
METHODS: The target model is the multitask convolutional neural network for information extraction from cancer pathology reports, where the data for training the model are from multiple state population-based cancer registries. This study proposes the following schemes to collect vocabularies from the cancer pathology reports; (a) words appearing in multiple registries, and (b) words that have higher mutual information. We performed membership inference attacks on the models in high-performance computing environments.
RESULTS: The comparison outcomes suggest that the proposed vocabulary selection methods resulted in lower privacy vulnerability while maintaining the same level of clinical task performance.

Keywords: Privacy, privacy-preserving training, deep learning, natural language processing, cancer epidemiology, artificial intelligence

*Corresponding author: Hong-Jun Yoon, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN, 37830, USA. Tel.: +1 865 241 2626; E-mail: yoonh@ornl.gov.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) – based automatic information extraction is an actively researched and developed topic because such tasks are massively labor-intensive, costly, and error prone [1,2]. Recent advances in deep learning (DL) have improved accuracy and reduced the burden of model training. One of the major advances of DL is that optimal features for prediction are constructed in the model building process, and manual curation of features is not needed. However, that desirable property of self-feature representation also poses the risk that information from the training dataset, either open or private, will be exposed or leaked [3,4]. Since DL models are agnostic to the nature of the information they contain, they cannot distinguish which portion of training data is open information or private information.

Open information is information that appears across the data samples in the corpus, which may contain key features that are correlated with the outcome and thus can be used in a classification model with robust performance beyond the dataset used to train the model. *Private information*, on the other hand, is highly specific to an individual observation, and includes but is not limited to some proper nouns or unique combinations of open information. In our application, we desired a model that maintains high predictive performance when transferred from a training set to a real-world production environment. Specifically, in the context of extracting information from an electronic cancer pathology report (e-path report) corpus, we developed a multi-task convolutional neural network (MT-CNN) model to identify primary cancer sites and their properties from unstructured text in pathology reports, and both types of information are present in the training set [5]. The training data includes open information such as words and phrases with characteristics of primary cancer sites (e.g., breast cancer, HER2) or histology (e.g., adenocarcinoma). Nevertheless, if some of the words and phrases are related to individual patients (e.g., name or other Health Insurance Portability and Accountability (HIPAA) protected health information (PHI)) or the specific name of the pathology laboratory, they are private information. Such private information is not useful in the model because it is unlikely to generalize to predictions on new corpora and should not be contributed to the information extraction tasks. However, there is a possibility that a ML/DL training algorithm will leverage private information to increase its accuracy. The process by which ML/DL training leverages private information to boost classification accuracy is called *overfitting*. Overfitting is observed when the accuracy of the training data is much larger than the prediction accuracy of the test data. Thus, overfitting increases accuracy of training samples, but it does not improve the accuracy of test data samples that have not been exposed to the model and may make it worse by decreasing generalization of the features. Although there are several techniques to avoid overfitting, there is no guarantee that private information will not be used as a signal in model predictions.

The underlying security vulnerability can be articulated as the hypothesis that the model may make confident decisions on samples already exposed for training, but it provides less confidence for testing samples. In other words, DL models behave differently with training data compared with data they have not seen. In generic DL models, those kinds of confidence and uncertainty can be obtained from the softmax layer output. By using distributions of softmax output from the model from the training samples (prior distributions), as well as distributions of output based on samples that have never been exposed to the model (posterior distributions), we can estimate the membership and association of the samples. This process is called a *membership inference attack* (MIA). However, it is impossible to obtain posterior distributions without the training samples, and if the entire set of training samples is available, then such attacks are not even needed. Shokri et al. [3] introduced a novel method of estimating such posterior distributions from a shadow model with a training dataset that possesses similar properties to those of the target model. We applied Shokri's approach to a cancer pathology report corpus. The algorithm estimates the posterior distribution of the association of data samples by using multiple shadow models. With simulated adversarial attacks, we can quantify how vulnerable any ML/DL model is to MIAs.

One intuitive way to avoid overfitting DL models for text data comprehension is to eliminate the noninformative words and tokens from the vocabulary set. However, there are no straightforward methods for accomplishing this. To address this gap, we proposed two approaches for optimal vocabulary selection. We then experimentally validated the effectiveness of the proposed approaches to the DL models for text comprehension and information extraction from cancer pathology reports provided by the seven participating cancer registries. Finally, we quantified the privacy vulnerability with the MIA accuracy scores as well as the clinical task performance scores.

This paper is organized as follows: in Section 2 we discuss related work and describe the research problem. In Section 3, we present the methods, including a brief description of the MIA algorithm, the cancer pathology report datasets, a target MT-CNN model used to perform the privacy attacks, proposed vocabulary selection approaches, and the study design with details of the simulation experiments and defined metrics. In Section 4 we present results of the study. Finally, in Sections 5 and 6 we present the discussion and conclusion.

2. Privacy attacks toward machine learning models related work

Homer et al. [6] studied MIAs statistically to estimate the probability that a particular record was part of a dataset; they documented attacks on the privacy of biological data. They demonstrated that it is possible to identify whether an individual participated in a study by knowing parts of the individual's genome and the summary statistics of the genome-based study. More recent studies have been published by Dwork et al. [7] and Backes et al. [8]. Dwork's research question focused on estimating how many adaptive chosen statistical queries can be correctly answered using samples from a distribution [7]. Backes demonstrated that MIAs also threaten the privacy of individuals who contribute their microRNA expressions to scientific studies using disease-specific datasets by relying only on published mean statistics [8]. In their suggested solution approaches, both Dwork and Backes used differential privacy [9] to decrease the probability of MIAs.

In [3], Shokri et al. studied MIA for ML as the attempt to discern the data in a query as part of the training dataset (i.e., to use the model to learn whether a particular record was part of the training data or not). ML presents what is known as a *black box setting*. That is, none of the following are available: knowledge of the model parameters, direct knowledge access to the algorithm's implementation, knowledge of the data distribution, or knowledge of all of the features in the dataset and/or the trained data. Thus, how can attackers perform an MIA? They can do so by using the intuition of overfitting, knowing that ML behaves differently with respect to the training data compared with data that the model has not seen.

Shokri et al. motivated their approach starting from the typical behavior in ML models: first, a query is submitted from the training set to the prediction API, and then a classification result vector R1 is received. Suppose an input that is not from the training set, but from outside the training set, can be submitted to the prediction API; in that case, another classification result vector, R2, will be received. A would-be attacker would need only to identify the difference between the two resulting classification responses: R1 and R2. ML is used to identify these differences; thus, an ML model can be trained to recognize the difference, which is known as an *attack model*. The attack model is basically a binary classifier: it observes a prediction vector and outputs the probability that this prediction vector is coming from the members versus the nonmembers.

The main question that Shokri's work answered is how to train an attack model without access to the training data. Shokri asserts that the main goal of an attack model is to learn the behavior of the target model using training data and to compare it with the behavior it exhibits toward data from outside the training set. The way to train the attack model without access to the training set is to learn the behavior not from the target model, but from other models that Shokri called *shadow models*. Assuming that the architecture of the target model is known, and that the attacker has some data that are in the same underlying distribution as the training data of the target model, then other models such as the shadow models can be trained. The attack model can be trained on the predictions that these shadow models produce on their training data versus their test data. Consequently, if the architecture of a shadow model is the same type as the architecture of the target model, and if the data used to train the shadow model are similar to the data used to train the target model (although with different parameters), then the shadow model will behave similarly to the target model with regard to the respective sets of testing and training data.

Other approaches have been used to study and prevent MIAs. In [10], Li et al. presented an adversarial training framework, the "Deepobfuscator," that prevented extracted features from being used to reconstruct raw images and infer private attributes, while the common data were used to train for image classification. In [11], Wang et al. proposed a sparse method based on single value decomposition to decrease parameters in CNNs to compress remote sensing images and thus protect from attacks against privacy. Chamikara et al. presented PABIDOT in [12], a nonreversible perturbation algorithm for privacy preservation of big data that uses optimal geometric transformations. Hao et al. [13] proposed a privacy-enhanced federated learning scheme for industrial AI. Shen et al. [14] proposed a morphed learning approach to deliver DL data efficiently and

securely. Jia et al. [15] proposed MemGuard, which incorporates a noise vector to a confidence score vector to turn it into an adversarial example that misleads the attacker's classifier. Chen et al. [16] proposed an MIA that leverages different outputs of an ML model by introducing the concept of machine unlearning (i.e., removal of requested data from the training dataset by the ML model owner). His work indicated that ML can have counterproductive effects on privacy. Song et al. [17] proposed to benchmark membership inference privacy risks by improving existing non-neural network-based inference attacks and proposed a new inference attack method based on a modification of prediction entropy. They also proposed benchmarks for defense mechanisms by accounting for adaptive adversaries with knowledge of the defense, and accounting for the trade-off between model accuracy and privacy risks. Song's benchmark attacks demonstrated that existing defense approaches are not as effective as had previously been reported. Song introduced a new approach to fine-grained privacy analysis by formulating and deriving a new metric called "privacy risk score" that measures an individual sample's likelihood of being a training member.

Compliance with privacy regulations to prevent attacks has also gained interest, and many approaches have been developed to suggest solutions [18–24]. Challenges and opportunities for genomics data sharing is explored in [25]. Abouelmehdi et al. surveyed security and privacy challenges in big data applied to healthcare [26] by Bonomi et al. Legal and ethical challenges to patient privacy in big data were explored in [27,28].

3. Methods

3.1. Membership inference attacks

The objective of MIA is to estimate if a certain data sample is associated with the training dataset. In the information extraction model for the e-path reports case, using this kind of attack allows us to identify whether a particular person is a cancer patient/survivor, which is a disclosure of very private health information and a serious lack of conformance with privacy policies.

However, the ML/DL model is a kind of black box (i.e., the model is not designed to expose the details of the inference process). To estimate the prior (identify whether a particular person is a cancer patient/survivor) from the posterior (inference output of the black box), a pair of the training corpus (the e-path reports for training the model) and its model outputs are required, which is not available for most cases.

The privacy attack model [3] is based upon the assumption that the posterior distribution can be estimated from a series of models trained by the same hyperparameters and by another set of training samples that may have a similar distribution to the training samples in the model to be attacked. We applied the following definitions and MIA approach:

1. M_{target} is the model that we were attempting to attack.
2. The attack attempts were trained by the set of training samples x_{target}.
3. In addition, we developed a series of shadow models, M^i_{shadow}, $i = 1 \cdots N$, that resemble the target model, where N is the number of shadow models.
4. For a given data corpus that had a similar distribution to the one for the target model, x_{shadow}, we provided N sets of training samples (shadow prior) to produce inferences (shadow posterior), y_{shadow}.
5. A variety of methods could be used to prepare the shadow dataset. In this study, we chose to apply a 50/50 random split to x_{shadow}: $x^{i,in}_{shodow}$ and $x^{i,out}_{shadow}$.
6. Then we trained the model, M^i_{shadow}, only with $x^{i,in}_{shadow}$, and collected the inferences from the model, $y^{i,in,c}_{shadow}$, $y^{i,out,c}_{shadow}$, $c = 1 \cdots C$, where C is the number of classes in the dataset.
7. Finally, we developed an attack model, M^c_{attack}, for a given $y^{i,in,c}_{shadow}$ and $y^{i,out,c}_{shadow}$, for $\forall i = 1 \cdots N$, in which we assigned the truth label, either 0 to $y^{i,out,c}_{shadow}$ or 1 to $y^{i,in,c}_{shadow}$.
8. C is the number of attack models that committed MIAs.

3.2. Dataset

The dataset for this study consisted of unstructured text in e-path reports from seven cancer registries: the California Cancer Registry, Kentucky Cancer Registry, Louisiana Tumor Registry, New Jersey State Cancer Registry, New Mexico Tumor Registry, Cancer Surveillance System (covering the Seattle-Puget Sound region), and Utah Cancer Registry. These registries are participants in the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) program. The study was executed in accordance with the institutional review board protocol DOE000619, approved by the Central Department of Energy Institutional Review Board on April 6, 2021 (initial approval on September 23, 2016).

We determined truth labels of the e-path reports based on the Cancer/Tumor/Case (CTC) database, which stores all diagnostic, staging, and treatment data for reportable neoplasms in the SEER Data Management System. Importantly, the CTC database was created using manual classification by trained and experienced cancer registry staff. We used the *International Classification of Diseases for Oncology* [29], Third Edition (ICD-O-3) coding convention for labeling the cases. The following five data fields and number of classes were used for the model: cancer site (70 classes), subsite (326 classes), laterality (7 classes), behavior (4 classes), and histology (639 classes).

3.3. MT-CNN model

The target model that we used to launch an MIA is the MT-CNN model [5], which is an extension of a CNN model for text classification [30] with additive multitask fully connected layers at the end. Training and optimization were based upon a gradient for each task. The summation of the validation loss determined the termination of the training from the dataset, and we reserved 10% of the training samples for this. The possibility of the model overfitting exists; not every task is equally easy or complicated. For example, the histology task had more than 500 class labels, but the behavior task had fewer than 10 labels. Also, there were some class labels with abundant training samples and some others with very few, mainly because of rare cancer sites or histologic types. Minor class labels made the MT-CNN model more complicated to train, degraded the task performance scores, and caused overfitting to the minor classes. Overfitting is well known to cause privacy vulnerability and is subject to MIAs.

3.4. Optimal vocabulary selections for privacy-preserving DL model training

The intrinsic idea of the privacy-preserving model training for natural language text comprehension is to identify which words are useful to the clinical tasks we are interested in and which are not and then to employ the useful words as the model's vocabulary. This approach is based on the understanding that words that may contain personally identifiable information (PII) and other PHI are not helpful to the clinical tasks. Even worse, those noninformative words may incur overfitting if those keywords are involved in the classification tasks.

Our previous study [31] trained the MT-CNN model with a vocabulary limited to the words and tokens that appeared only in the description field of the concept unique identifiers in the Unified Medical Language Systems. This was based upon the assumption that such a text corpus should not include any patient-specific keywords (e.g., patient names). However, there is no guarantee that the Unified Medical Language System's vocabulary only contains keywords and tokens for open information, so the possibility of privacy leakage persists.

In this study, we propose the following two approaches of eliminating the private information by identifying noninformative keywords in the model training.

3.4.1. Words appearing across multiple cancer registries – intersection approach

The underlying hypothesis is that if the words appear across the multiple cancer registries, those may represent a commonality – information and characteristics that describe what the cancer pathology reports should possess. On the contrary, if any of the keywords are used only by a specific cancer registry, such words may have limited knowledge or no information on the clinical tasks.

We performed the training of MT-CNN models with keywords that have $S(word) \geqslant S_{threshold}$, and discard $S(word) < S_{threshold}$, where $S(word)$ is the number of registries the keyword *word* appears in the e-path reports. Note, technically, $S_{threshold}$ cannot be greater than the number of cancer registries in the training set. Therefore, in our experiments, $\max(S_{threshold}) = 6$. Consequently, the total number of keywords in the vocabulary set is subject to the $S_{threshold}$. For a given set of training corpus, it is evident the vocabulary size is decreased if $S_{threshold}$ increased.

We are interested in observing if the MT-CNN model with the vocabulary with higher $S_{threshold}$

1. maintains the same clinical task scores, and
2. reduces privacy vulnerability in MIA accuracy.

3.4.2. Words possess useful information for the given clinical tasks – mutual information approach

The first strategy is simple and intuitive. However, it is applicable only if the data corpus was composed by the subset of data from multiple data sources, and the properties of the subset of data should be identical (i.e., multiple cancer registries in our study). As a counterpart, we propose a measure of the usefulness of the keywords based on the mutual information (MI).

In this study, we calculated MI by treating each keyword and class label as a binary random variable

(present, or not present). For a given keyword and class combination, the MI was normalized according to the entropy of the class. To score the keywords, the maximum value of the normalized MI was calculated across all classes and tasks. The top N keywords, according to their score, were then retained for model training. Therefore, the total number of keywords in the vocabulary of the MT-CNN model becomes controllable. However, the optimal number of N needs to be determined by experiments. Our experiments are intended to observe

1. how the task performance scores and privacy vulnerability vary if the number of keywords N changes, and
2. to what extent does the MI approach award the same level of clinical task performance and MIA accuracy scores compared to the intersection approach with a similar number of keywords in the vocabulary set.

3.5. Study design

The primary purpose of the study is to assess if the privacy-preserving vocabulary selection approaches we introduced in Sections 3.4.1 and 3.4.2 are effective means of protection against MIAs. We quantified the vulnerability of the MT-CNN models for text comprehension of the e-path reports that have a chance of including private information. If there is private information in the model, some of those few outliers may drive classification and decision-making, which will increase the chance of successful MIAs.

We designed a series of privacy-attacking experiments on the models trained with data from the seven cancer registries, where the model (target model) was trained with the data from six registries (target model), then the remaining one registry performed MIAs on the target model.

We developed multiple shadow models through random sampling to obtain reliable posterior distributions, and we developed one attack model for each class label. We designed MIAs against the cancer main site classification task of the MT-CNN, which consisted of $C = 70$ class labels. Consequently, in this study, we

1. trained a MT-CNN model with the data from six registries, (target model M_{target});
2. developed 100 shadow models (M^i_{shadow}, $i = 1 \cdots 100$) using the same MT-CNN model architecture with the bootstrap sampled data from the remaining registry;
3. implemented 70 attack models M^c_{attack} with the outputs from the 100 shadow models;
4. performed MIAs with the attack models to the target model;
5. determined the classification accuracy of the e-path reports identified (if it belongs to the training corpus [i.e., data from the six registries) of the target model or not (the one from the remaining registry); and
6. repeated the experiment seven times.

3.6. Performance measure

The success of MIAs depends on whether the attack model, M^c_{attack}, classifies correctly if a document (in this context, an e-path report) is included in the training set by observing the softmax output of the document from the target model, M_{target}. In other words, the attack model is a two-class classifier estimating if a report is inside or outside the training set. This study employed simple accuracy as a performance measure, where the chance level is 0.5, and the perfect classification equaled 1.0. Note, the number of data samples between the training samples inside or outside the training set may not be identical; it depends on the number of available samples in the target and attack registries.

Experiments were performed on three models: (1) the MT-CNN model with vocabulary of words and tokens that appeared at least five times in the corpus (baseline); (2) the model with vocabulary selected by their appearance across multiple cancer registries (intersection approach); and (3) the model with vocabulary selected by the utility scores determined by the MI (MI approach). Results of the experiments are the quantification of privacy vulnerability in terms of the MIAs, as well as the clinical task performance.

The clinical task performance scores could be important indicators if trade-offs between the task performance and data privacy exist. We employed both micro- and macro-averaged F1 scores because the tasks possess severe class imbalance. The micro-averaged F1 scores are useful if we observe the task performance for each document, whereas the macro-averaged F1 scores weigh equally across class labels, thus reflecting impact by the class imbalance.

Note, we are training classifiers of multitask learning mechanisms and processing five tasks, resulting in 10 F1 scores; consequently, comparing clinical task performances between the models becomes utterly complicated. To mitigate the complication, we introduced averaged F1 scores. The averaged F1 scores may not have clinical implication, but they are listed for ease of comparison.

Table 1
Privacy vulnerability quantified by the accuracy scores of the MIAs on electronic cancer registry data. The MIA is binary decision, where 1.0 is the perfect identification, and 0.5 is the chance level. In the column index, B is toward the baseline model, I2, I3, I4, I5, I6 is to the models with $S_{threshold} = \{2, 3, 4, 5, 6\}$, and MI to the models with the vocabulary selected by the higher mutual information. The numbers 1, 10, 20, 40, 60, 80, and 100 on the left column indicate the number of shadow models applied to the MIAs. For the deeper understanding from the experiments, we divided the attacks toward the class labels that have available number of average training samples of < 100, $\geqslant 100$ and < 1000, $\geqslant 1000$ and < 10000, and $\geqslant 10000$

		Intersection					Mutual information			
	B	I2	I3	I4	I5	I6	MI	MI	MI	MI
# shadow models	58322	57860	35396	25026	18274	12529	20000	10000	5000	2000
All										
1	0.560	0.587	0.559	0.538	0.543	0.542	0.548	0.530	0.532	0.513
10	0.571	0.605	0.572	0.547	0.557	0.554	0.561	0.539	0.537	0.520
20	0.574	0.607	0.576	0.551	0.562	0.557	0.562	0.543	0.541	0.523
40	0.577	0.608	0.578	0.554	0.559	0.560	0.563	0.543	0.541	0.523
60	0.579	0.611	0.577	0.555	0.560	0.559	0.564	0.544	0.543	0.525
80	0.578	0.611	0.578	0.555	0.561	0.557	0.563	0.543	0.541	0.524
100	0.578	0.610	0.579	0.554	0.560	0.559	0.564	0.543	0.542	0.523
< 100										
1	0.623	0.692	0.593	0.586	0.584	0.593	0.621	0.567	0.594	0.555
10	0.651	0.721	0.644	0.639	0.643	0.637	0.681	0.607	0.595	0.567
20	0.666	0.747	0.682	0.642	0.667	0.651	0.675	0.642	0.630	0.593
40	0.687	0.755	0.689	0.660	0.639	0.668	0.676	0.633	0.624	0.591
60	0.693	0.785	0.694	0.674	0.639	0.641	0.681	0.615	0.629	0.608
80	0.695	0.752	0.682	0.674	0.662	0.638	0.664	0.620	0.623	0.614
100	0.670	0.762	0.681	0.659	0.641	0.650	0.659	0.627	0.631	0.590
$>= 100$ and < 1000										
1	0.571	0.601	0.571	0.545	0.549	0.549	0.556	0.535	0.532	0.514
10	0.581	0.616	0.578	0.548	0.558	0.556	0.565	0.537	0.539	0.519
20	0.581	0.617	0.581	0.550	0.563	0.562	0.564	0.541	0.539	0.518
40	0.586	0.620	0.585	0.553	0.566	0.564	0.566	0.542	0.543	0.519
60	0.585	0.620	0.584	0.552	0.565	0.567	0.567	0.544	0.544	0.520
80	0.585	0.623	0.584	0.553	0.566	0.563	0.570	0.543	0.541	0.519
100	0.585	0.622	0.587	0.553	0.567	0.561	0.569	0.542	0.542	0.521
$>= 1000$ and < 10000										
1	0.536	0.555	0.538	0.521	0.526	0.525	0.528	0.516	0.515	0.506
10	0.542	0.561	0.543	0.525	0.531	0.530	0.532	0.519	0.516	0.508
20	0.543	0.561	0.545	0.527	0.532	0.530	0.534	0.520	0.518	0.508
40	0.544	0.561	0.545	0.529	0.530	0.531	0.533	0.520	0.516	0.509
60	0.545	0.560	0.543	0.527	0.532	0.531	0.533	0.521	0.517	0.509
80	0.544	0.562	0.543	0.529	0.531	0.532	0.533	0.521	0.518	0.508
100	0.545	0.563	0.544	0.529	0.530	0.528	0.533	0.522	0.517	0.507
$>= 10000$										
1	0.515	0.511	0.517	0.513	0.511	0.512	0.528	0.516	0.515	0.506
10	0.518	0.518	0.517	0.511	0.513	0.513	0.532	0.519	0.516	0.508
20	0.518	0.519	0.518	0.516	0.515	0.515	0.534	0.520	0.518	0.508
40	0.518	0.521	0.518	0.514	0.515	0.516	0.533	0.520	0.516	0.509
60	0.520	0.520	0.518	0.514	0.513	0.516	0.533	0.521	0.517	0.509
80	0.519	0.520	0.517	0.514	0.515	0.513	0.533	0.521	0.518	0.508
100	0.518	0.520	0.518	0.515	0.516	0.519	0.533	0.522	0.517	0.507

3.7. Experiments

The experiments training MT-CNN models for classifying cancer pathology reports were performed with the Keras [32] and TensorFlow [33] back end of the IBM Watson Machine Learning tool kits on the Summit supercomputer at the Oak Ridge Leadership Computing Facility. The models were trained in parallel with the Exascale Computing Project's Cancer Distributed Learning Environments [34].

Training the models required the vocabulary sets collected by the two approaches described in the Sections 3.4.1 and 3.4.2, along with the baseline method, which is the current practice of the training that we eliminated the underrepresented tokens that appear fewer than five times. In summary, the comparison

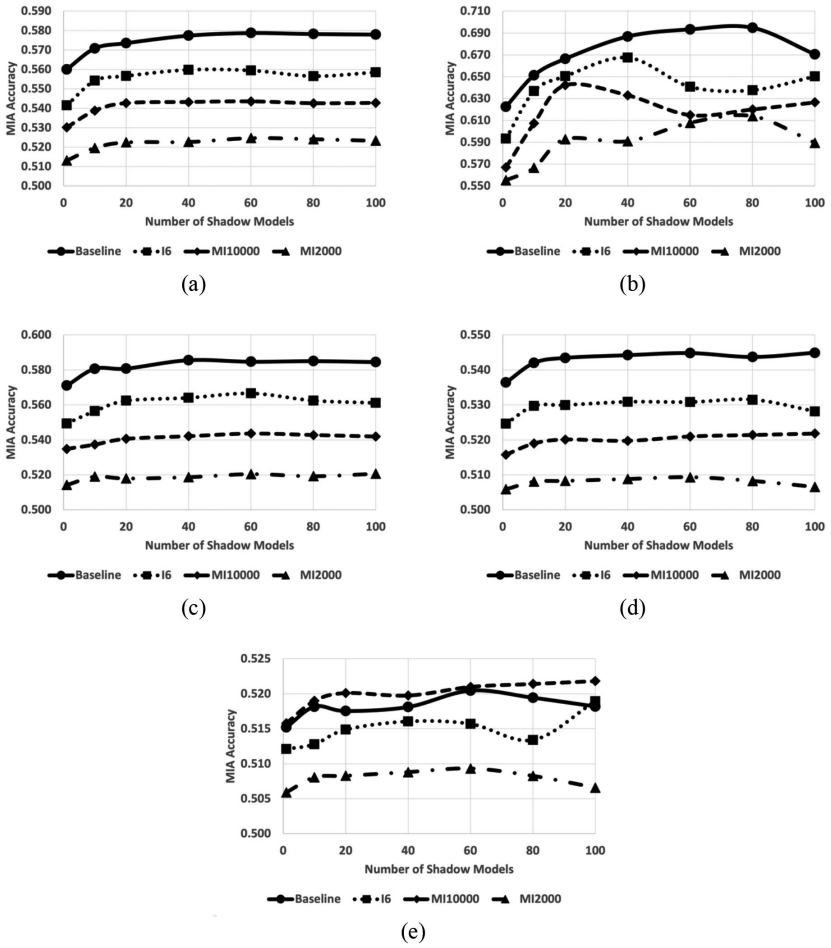

Fig. 1. Average MIA accuracy scores toward the MT-CNN models with vocabulary of baseline (circle), keywords appearing all six registries (I6, square), 10,000 keywords of highest mutual information (MI10000, diamond), and 2,000 keywords of highest mutual information (MI2000, triangle) across (a) all the attack models, (b) attack models to the class labels with average training samples < 100, (c) $\geqslant 100$ and < 1000, (d) $\geqslant 1000$ and < 10000, (e) $\geqslant 10000$.

used the following approaches: baseline, five intersection thresholds $S_{threshold} = \{2, 3, 4, 5, 6\}$, and mutual information with the following number of keywords: $N = \{20000, 10000, 5000, 2000\}$.

4. Results

4.1. Privacy vulnerability

Table 1 lists the MIA accuracy scores to the MT-CNN models with the proposed optimal vocabulary selection approaches, with respect to the number of the shadow models. In accordance with [3], the MIA accuracy increased as we applied more shadow models to the simulation. However, the accuracy did not improve if more than 60 shadow models were applied.

The 70 class labels of site labels were divided by the prevalence of data samples. Table 1 lists the MIA accuracy separately by their availability of training samples, which resulted in higher MIA accuracy toward the class labels that have fewer than 100 training samples available, although the MIA accuracy toward the highly prevalent cancer types ($\geqslant$ 10,000) was nearly chance level. The observation supported our hypothesis that the overfitting of the DL models could be happening more on the underrepresented labels because there was a higher chance of successful MIAs toward the minor classes. The intersection approach was effective. It reduced the MIA accuracy from 0.578 to 559 for all the cancer site labels. The mutual information approach allowed an even lower MIA accuracy of 0.543 (M10000 model) and 0.523 (M2000 model).

Figure 1 illustrates several interesting observations

from the experiments. We observed that the models exposed lower security vulnerability as the algorithm secured vocabularies more tightly. On average, the baseline model has a vocabulary of 58,332 words and tokens, the I6 models used 12,529 words, the MI1000 models used 10,000 words, and the MI2000 models used 2,000 keywords. Overall, the baseline models were more vulnerable to MIAs. The MI2000 models were the most secure across the models tested in the experiments. Class labels with fewer training samples were more vulnerable to the MIA. The MIA accuracy scores in Fig. 1b were the highest, which is toward the set of class labels. The MIA accuracy scores in Fig. 1e, with the most prevalent class labels, were the lowest. The MIA accuracy scores remain stable as we applied more than 50 shadow models except for the ones with the least prevalent labels and the most prevalent labels. Presumably, the variations in Fig. 1b were due to the small number of available samples, whereas the fluctuations in Fig. 1e were due to the low MIA accuracy scores.

Note, the I2 model recorded higher MIA accuracy than the baseline method even though the vocabulary numbers are similar (58,332 vs. 57,860). The baseline model discarded the keywords with a frequency less than five, but the I2 model retained some of the low-frequency keywords. Those keywords and terms may incur overfitting to the training samples in the minority classes, resulting in the success of the MIA.

4.2. Clinical task performance

The clinical task performance scores in F1 metric, as well as the average scores, are shown in Table 2. The models with intersection approach recorded the competitive F1 scores with the baseline model in micro- and macro-averaged F1 scores. (Micro F1: $\sim$ 0.852–0.855, Macro F1: $\sim$ 0.543–0.570). On the contrary, the models with MI approach showed a performance decrease if we applied a few keywords. The MI2000 model recorded 0.747 in the micro-F1 score and 0.413 in the macro-F1. Although the MI2000 model was the most secure model against the MIA (Table 1), it sacrificed clinical task performance to achieve privacy.

Figure 2, which plotted the F1 scores as a function of the number of words and tokens in the vocabulary, visualizes the trend of the task performance scores. Overall, the intersection approaches hold the same accuracy scores as the scores from the baseline model. Variability across the intersection-based models was observed to some extent, but it is not evident that it is subject to the size of vocabularies. On the contrary, the F1 scores dropped drastically when the vocabulary size was smaller than 10,000.

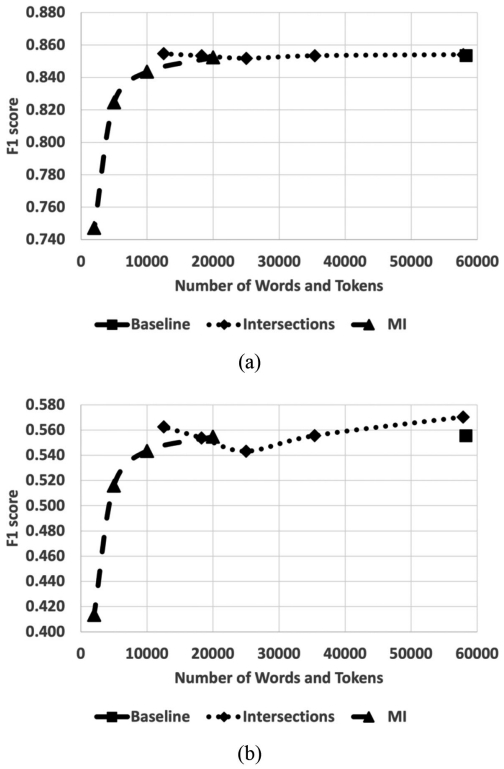

Fig. 2. Average of clinical task performance scores both in (a) micro-averaged and (b) macro-averaged F1 scores. We averaged all five information extraction tasks: site, subsite, laterality, histology, and behavior.

5. Discussion

The MIA with multiple shadow models is based on the following two assumptions. First, the adversary learns about the target model architecture and hyperparameters to train. The idea is that the adversary can develop multiple shadow models that mimic the target model's behavior, thus estimating the posterior distribution and identifying whether it is the output from the samples included in the training corpus or samples that have never been exposed to the model before. The target model architecture and its hyperparameters may be gleaned from the relevant research papers from the developers, but not thoroughly. Second, the adversary should have a good volume of quality data samples to train the faithful shadow models to mimic the target model. Such abundant data may not be easily acquired for biomedical informatics.

The MIA settings and data in this study would be

Table 2

Clinical task performance scores in micro- and macro-average F1 scores of the five information extraction tasks: site, subsite, laterality, histology, and behavior. The R on the left column denotes the attack registry that the data reserved for performing MIAs, A is the approaches; B (baseline), I2, I3, I4, I5, I6 for $S_{threshold}$, and MI, # denote the number of keywords in the vocabulary of the MT-CNN classifier. The μ on the last row is the numerical average across the above rows

			Site		Subsite		Laterality		Histology		Behavior		Average	
R	A	#	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
1	B	63121	0.925	0.690	0.683	0.360	0.914	0.528	0.778	0.352	0.974	0.860	0.855	0.558
	I2	62732	0.926	0.701	0.683	0.354	0.917	0.533	0.782	0.368	0.975	0.910	0.856	0.573
	I3	38519	0.926	0.691	0.683	0.342	0.915	0.501	0.778	0.336	0.975	0.912	0.855	0.556
	I4	27411	0.926	0.696	0.681	0.354	0.915	0.522	0.779	0.342	0.973	0.884	0.855	0.560
	I5	20201	0.926	0.704	0.683	0.355	0.918	0.537	0.781	0.340	0.974	0.909	0.856	0.569
	I6	14088	0.927	0.698	0.683	0.347	0.915	0.532	0.781	0.356	0.975	0.910	0.856	0.569
	MI	20000	0.924	0.697	0.683	0.346	0.917	0.548	0.782	0.347	0.975	0.913	0.856	0.570
	MI	10000	0.924	0.677	0.642	0.344	0.912	0.516	0.770	0.330	0.972	0.896	0.844	0.553
	MI	5000	0.916	0.664	0.592	0.321	0.907	0.493	0.754	0.320	0.963	0.848	0.827	0.529
	MI	2000	0.889	0.591	0.533	0.221	0.691	0.257	0.719	0.258	0.922	0.749	0.751	0.415
2	B	52495	0.919	0.659	0.658	0.308	0.911	0.491	0.770	0.303	0.976	0.855	0.847	0.523
	I2	51993	0.919	0.682	0.667	0.362	0.911	0.509	0.773	0.399	0.973	0.860	0.849	0.562
	I3	32310	0.920	0.658	0.661	0.300	0.912	0.504	0.768	0.307	0.976	0.818	0.847	0.517
	I4	23151	0.919	0.653	0.656	0.307	0.913	0.513	0.767	0.302	0.975	0.851	0.846	0.525
	I5	17105	0.923	0.679	0.669	0.340	0.913	0.519	0.774	0.386	0.977	0.884	0.851	0.562
	I6	11876	0.922	0.684	0.669	0.345	0.913	0.502	0.774	0.372	0.977	0.870	0.851	0.555
	MI	20000	0.917	0.653	0.652	0.301	0.912	0.520	0.768	0.299	0.975	0.852	0.845	0.525
	MI	10000	0.917	0.648	0.627	0.298	0.911	0.483	0.768	0.289	0.975	0.856	0.839	0.515
	MI	5000	0.909	0.641	0.583	0.283	0.898	0.490	0.743	0.287	0.961	0.781	0.819	0.497
	MI	2000	0.881	0.592	0.533	0.232	0.689	0.276	0.704	0.277	0.926	0.743	0.747	0.424
3	B	61834	0.925	0.688	0.678	0.348	0.913	0.522	0.778	0.362	0.976	0.892	0.854	0.562
	I2	61492	0.926	0.686	0.681	0.350	0.914	0.530	0.780	0.374	0.976	0.907	0.855	0.569
	I3	37650	0.925	0.681	0.680	0.345	0.915	0.541	0.781	0.339	0.976	0.903	0.855	0.562
	I4	26470	0.922	0.640	0.665	0.293	0.914	0.500	0.768	0.247	0.976	0.886	0.849	0.513
	I5	19046	0.925	0.683	0.680	0.335	0.915	0.520	0.779	0.346	0.974	0.897	0.855	0.556
	I6	12734	0.925	0.676	0.679	0.355	0.914	0.543	0.782	0.351	0.976	0.910	0.855	0.567
	MI	20000	0.926	0.682	0.680	0.349	0.914	0.523	0.780	0.368	0.975	0.902	0.855	0.565
	MI	10000	0.924	0.674	0.643	0.325	0.910	0.514	0.772	0.297	0.975	0.863	0.845	0.535
	MI	5000	0.915	0.654	0.597	0.297	0.905	0.499	0.748	0.307	0.963	0.835	0.826	0.518
	MI	2000	0.889	0.584	0.536	0.249	0.696	0.303	0.717	0.288	0.920	0.741	0.752	0.433
4	B	62129	0.924	0.672	0.675	0.343	0.915	0.524	0.777	0.358	0.976	0.902	0.853	0.560
	I2	61711	0.924	0.671	0.677	0.360	0.914	0.537	0.779	0.376	0.976	0.923	0.854	0.574
	I3	37678	0.924	0.674	0.677	0.355	0.913	0.531	0.777	0.375	0.975	0.915	0.853	0.570
	I4	26610	0.924	0.680	0.675	0.332	0.914	0.527	0.774	0.342	0.975	0.903	0.853	0.557
	I5	19335	0.925	0.673	0.674	0.349	0.917	0.530	0.779	0.378	0.977	0.916	0.854	0.569
	I6	13065	0.925	0.671	0.678	0.350	0.917	0.534	0.777	0.351	0.977	0.905	0.855	0.562
	MI	20000	0.924	0.684	0.679	0.361	0.915	0.524	0.780	0.378	0.977	0.910	0.855	0.571
	MI	10000	0.924	0.677	0.643	0.350	0.910	0.527	0.771	0.357	0.976	0.901	0.845	0.562
	MI	5000	0.916	0.644	0.599	0.302	0.907	0.503	0.749	0.316	0.967	0.858	0.827	0.525
	MI	2000	0.892	0.593	0.546	0.223	0.699	0.326	0.714	0.243	0.924	0.754	0.755	0.428
5	B	55644	0.928	0.673	0.683	0.350	0.915	0.540	0.780	0.354	0.974	0.887	0.856	0.561
	I2	54366	0.926	0.679	0.686	0.371	0.913	0.538	0.781	0.376	0.973	0.914	0.856	0.575
	I3	33098	0.927	0.670	0.683	0.362	0.915	0.544	0.781	0.388	0.971	0.895	0.856	0.572
	I4	23393	0.928	0.677	0.684	0.344	0.915	0.545	0.783	0.374	0.975	0.929	0.857	0.574
	I5	17200	0.929	0.679	0.679	0.343	0.914	0.533	0.780	0.356	0.975	0.909	0.855	0.564
	I6	11898	0.928	0.677	0.681	0.349	0.916	0.531	0.778	0.334	0.973	0.924	0.855	0.563
	MI	20000	0.928	0.681	0.682	0.383	0.915	0.546	0.784	0.392	0.974	0.907	0.857	0.582
	MI	10000	0.927	0.671	0.640	0.312	0.911	0.528	0.773	0.302	0.974	0.900	0.845	0.543
	MI	5000	0.914	0.634	0.594	0.288	0.903	0.503	0.749	0.259	0.964	0.810	0.825	0.499
	MI	2000	0.882	0.572	0.521	0.219	0.690	0.292	0.704	0.262	0.915	0.734	0.742	0.416
6	B	58310	0.926	0.685	0.681	0.348	0.913	0.521	0.777	0.340	0.975	0.894	0.854	0.557
	I2	57998	0.923	0.682	0.679	0.358	0.911	0.541	0.777	0.388	0.972	0.889	0.852	0.572
	I3	34846	0.924	0.675	0.681	0.359	0.912	0.540	0.779	0.377	0.974	0.917	0.854	0.574
	I4	24308	0.927	0.690	0.683	0.354	0.915	0.540	0.780	0.356	0.975	0.892	0.856	0.566
	I5	17623	0.926	0.678	0.680	0.342	0.915	0.529	0.776	0.325	0.976	0.916	0.854	0.558
	I6	12086	0.926	0.677	0.677	0.359	0.914	0.530	0.778	0.350	0.974	0.897	0.854	0.563

Table 2, continued

			Site		Subsite		Laterality		Histology		Behavior		Average	
R	A	#	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
	MI	20000	0.925	0.709	0.681	0.372	0.913	0.531	0.779	0.377	0.974	0.885	0.855	0.575
	MI	10000	0.923	0.667	0.628	0.321	0.912	0.517	0.770	0.295	0.973	0.904	0.841	0.541
	MI	5000	0.914	0.658	0.589	0.299	0.902	0.511	0.751	0.307	0.960	0.830	0.823	0.521
	MI	2000	0.883	0.559	0.523	0.217	0.692	0.309	0.700	0.215	0.914	0.732	0.742	0.406
7	B	54724	0.924	0.675	0.679	0.350	0.916	0.540	0.781	0.365	0.974	0.898	0.855	0.566
	I2	54728	0.926	0.686	0.678	0.363	0.919	0.531	0.783	0.353	0.975	0.903	0.856	0.567
	I3	33673	0.923	0.657	0.674	0.327	0.917	0.515	0.778	0.300	0.975	0.892	0.853	0.538
	I4	23841	0.918	0.630	0.659	0.282	0.916	0.507	0.765	0.232	0.976	0.890	0.847	0.508
	I5	17409	0.920	0.632	0.656	0.279	0.913	0.483	0.771	0.246	0.975	0.853	0.847	0.498
	I6	11956	0.925	0.676	0.682	0.349	0.917	0.541	0.783	0.336	0.976	0.893	0.856	0.559
	MI	20000	0.918	0.633	0.652	0.281	0.915	0.508	0.769	0.239	0.970	0.813	0.844	0.495
	MI	10000	0.923	0.687	0.644	0.353	0.916	0.525	0.776	0.350	0.971	0.865	0.846	0.556
	MI	5000	0.914	0.660	0.593	0.313	0.908	0.516	0.751	0.318	0.964	0.811	0.826	0.523
	MI	2000	0.877	0.541	0.517	0.182	0.696	0.307	0.695	0.176	0.921	0.653	0.741	0.372
μ	B	58322	0.924	0.678	0.677	0.344	0.914	0.524	0.777	0.348	0.975	0.884	0.853	0.555
	I2	57860	0.924	0.684	0.679	0.360	0.914	0.531	0.779	0.376	0.974	0.901	0.854	0.570
	I3	35396	0.924	0.672	0.677	0.341	0.914	0.525	0.778	0.346	0.975	0.893	0.853	0.556
	I4	25026	0.923	0.666	0.672	0.324	0.915	0.522	0.774	0.313	0.975	0.891	0.852	0.543
	I5	18274	0.925	0.675	0.674	0.335	0.915	0.522	0.777	0.340	0.975	0.898	0.853	0.554
	I6	12529	0.925	0.680	0.678	0.351	0.915	0.530	0.779	0.350	0.975	0.901	0.855	0.562
	MI	20000	0.923	0.677	0.673	0.342	0.914	0.529	0.778	0.343	0.974	0.883	0.852	0.555
	MI	10000	0.923	0.671	0.638	0.329	0.912	0.516	0.771	0.317	0.974	0.884	0.844	0.543
	MI	5000	0.914	0.651	0.592	0.300	0.904	0.502	0.749	0.302	0.963	0.825	0.825	0.516
	MI	2000	0.885	0.576	0.530	0.220	0.693	0.296	0.708	0.245	0.920	0.729	0.747	0.413

Note: Average in the right column is the numerical average of micro- and macro-F1 scores of all five tasks. Those quantities may not carry any clinical implication, but we introduced this for easier review and comparison across models.

the best-case scenario for attackers (and the worst case for the cancer registries). In this study, the adversary knows the model architecture very well, possessed the real-world cancer data samples to develop shadow models, and gained access to high-performance computing resources to develop many shadow models in parallel. For those reasons, the MIA accuracy scores reported by these experiments may not measure the real-world threat in the field, but nonetheless, they help quantify the relative privacy vulnerability across the models with various vocabulary selection strategies.

5.1. Privacy vulnerability

The easiest way to avoid privacy threats of MIA on the DL models is to increase the volume of the training data corpus. As shown in Table 1, the MIA accuracy decreased drastically compared to the class labels trained by many cancer pathology reports. However, increasing the size of the training corpus is not always trivial for some domains and tasks, and this is true for biomedical and clinical data. If the tasks are for rare diseases and the data samples are not easily augmented or simulated, such as clinical text data, there will be only a limited volume of relevant training corpus.

Curation of the keywords in the vocabulary of the DL models helps reduce the privacy threats. We showed that the critical factor of the successful MIA attacks could be due to the overfitting of the DL models. The way to avoid such phenomena is by collecting a vocabulary set with informative keywords for the clinical tasks and training the DL models with it. In this paper, we proposed two approaches to collecting the informative keyword vocabulary set: intersection and mutual information. The MIA accuracy scores reported in Table 1 and the illustration of the accuracy trend in Fig. 1 demonstrated that both proposed approaches effectively reduced privacy vulnerability.

5.2. Task performance

Securing data and PII in a training corpus often decreases the task performance of the DL models, which may be inevitable. This study also observed that the most secure model (MI2000 model) recorded the lowest clinical task performance (Baseline: 0.853/0.555, MI2000: 0.747/0.413). The degradation was more severe according to the macro-averaged F1 scores, which suggests that the reduction of keywords could affect more than the underrepresented class labels.

Performance decreases may not be tolerable for some tasks, such as those that are mission critical. The inter-

section approach effectively positions the vocabulary set to achieve privacy-preserving DL model training, while maintaining the clinical task performance (Baseline: 0.853/0.555, Intersection6: 0.855/0.562). A vocabulary size of 12,529 words seems optimal for the mutual information approach too. The only disadvantage of the intersection approach is that it requires multiple data providers. In this study, this was achieved by the support of multiple cancer registries; however, this luxury is not always available.

The mutual information approach is also effective and flexible. With the proper choice of the vocabulary size, N, we can control the trade-offs between data security and task performance. A disadvantage is that this approach may require a series of trials to choose the optimal threshold.

Eliminating noninformative keywords and terms gives us another benefit in that we can conserve computing resources. In the DL models for processing natural language text, the majority of the memory is consumed by the word or token embedding matrix. By decreasing the informative keywords in the vocabulary, the DL model can allocate less memory for the embedding layer and spend less time to update the embedding vectors, thus producing more compact models while maintaining task performance.

6. Conclusions

This paper has identified a potential threat of privacy vulnerability via membership identification. Such a threat could be critical to the DL models trained by the clinical reports that may contain PHI and PII. We hypothesized that the membership identification is the characteristic of DL model training in that it determines the optimal features to maximize the accuracy. We developed a simulation of MIA to the MT-CNN for clinical text classification from e-path reports from the participating cancer registries, which quantifies the privacy vulnerability of the target model.

Our solution to the problem is the optimal selection of informative features. We proposed two approaches: intersection and mutual information. The former method takes keywords and tokens that appear across multiple cancer registries, based on the assumption that if the words are used by multiple registries, those common words are key components that define the report. The latter method is based on the utility score of the terms and tokens, which we quantified as mutual information in the study. Our study demonstrated that the proposed approaches decreased privacy vulnerability and maintained clinical task performance.

Acknowledgments

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 was also supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by DOE and the NCI 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 Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. This work also was supported by the Laboratory Directed Research and Development (LDRD) program of Oak Ridge National Laboratory, under LDRD project 9831.

The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco; contract HHSN261201800015I awarded to the University of Southern California; and contract HHSN261201800009I awarded to the Public Health Institute. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their contractors and subcontractors.

Kentucky Cancer Registry 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.

Louisiana Tumor Registry data were collected using funding from NCI and the Surveillance, Epidemiology and End Results (SEER) Program (HHSN26120180000 7I), the CDC's National Program of Cancer Registries (NPCR) (NU58DP006332-02-00) as well as the State of Louisiana.

New Jersey State Cancer Registry data were collected using funding from NCI and the Surveillance, Epidemi-

ology and End Results (SEER) Program (HHSN26120 1300021I), 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.

New Mexico Tumor Registry's participation in this project was supported by Contract HHSN26120180001 4I, Task Order HHSN26100001 from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) Program.

The Cancer Surveillance System is supported by the National Cancer Institute's SEER Program (Contract Award HHSN261291800004I) and with additional funds provided by the Fred Hutchinson Cancer Research Center.

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. NU58DP0063 200, with additional support from the University of Utah and Huntsman Cancer Foundation.

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

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).

Author contributions

Conception: Hong-Jun Yoon, Christopher Stanley, J. Blair Christian
Interpretation or analysis of data: Hong-Jun Yoon, Christopher Stanley, J. Blair Christian
Preparation of the manuscript: Hong-Jun Yoon, Hilda Klasky, Christopher Stanley, J. Blair Christian, Andrew Blanchard
Revision for important intellectual content: Eric Durbin, Xiao-Cheng Wu, Antoinette Stroup, Jennifer Doherty, Stephen Schwartz, Charles Wiggins, Mark Damesyn, Linda Coyle
Supervision: Georgia Tourassi

References

[1] J.X. Qiu, H. 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* **22**(1) (2017), 244–251.

[2] H. Yoon, J.P. Gounley, S. Gao, M. Alawad, A. Ramanathan and G.D. Tourassi, Model-based hyperparameter optimization of convolutional neural networks for information extraction from cancer pathology reports on HPC, in: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1–4.

[3] R. Shokri, M. Stronati, C. Song and V. Shmatikov, Membership inference attacks against machine learning models, in: *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[4] G. Ateniese, L.V. Mancini, A. Spognardi, A. Villani, D. Vitali and G. Felici, Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers, *International Journal of Security and Networks* **10**(3) (2015), 137–150.

[5] M. Alawad, S. Gao, J.X. Qiu, H. Yoon, J.B. Christian, L. Penberthy, B. Mumphrey, X. Wu, L. Coyle and G.D. 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* **27**(1) (2020), 89–98.

[6] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, S.F. Nelson and D.W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, *PLoS Genet* **4**(8) (2008), e1000167.

[7] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold and A.L. Roth, Preserving statistical validity in adaptive data analysis, in: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, 2015, pp. 117–126.

[8] M. Backes, P. Berrang, M. Humbert and P. Manoharan, Membership privacy in micro RNA-based studies, in: *Proceedings of the 2016 ACMSIGSAC Conference on Computer and Communications Security*, 2016, pp. 319–330.

[9] C. Dwork, Differential privacy: A survey of results, in: *International Conference on Theory and Applications of Models of Computation*, 2008, pp. 1–19.

[10] A. Li, J. Guo, H. Yang and Y. Chen, Deep obfuscator: Adversarial training framework for privacy-preserving image classification, *arXiv preprint arXiv:1909.04126*, 2019.

[11] E.K. Wang, N. Zhe, Y. Li, Z. Liang, X. Zhang, J. Yu and Y. Ye, A sparse deep learning model for privacy attack on remote sensing images, *Mathematical Biosciences and Engineering* **16**(3) (2019), 1300.

[12] M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe and I. Khalil, Efficient privacy preservation of big data for accurate data mining, *Information Sciences*, 2020, 527.

[13] M. Hao, H. Li, X. Luo, G. Xu, H. Yang and S. Liu, Efficient and privacy-enhanced federated learning for industrial artificial intelligence, *IEEE Transactions on Industrial Informatics* **16**(10) (2019), 6532–6542.

[14] J. Shen, J. Liu, Y. Chen and H. Li, Towards efficient and secure delivery of data for deep learning with privacy-preserving, *arXiv preprint arXiv:1909.07632*, 2019.
[15] J. Jia, A. Salem, M. Backes, Y. Zhang and N.Z. Gong, Memguard: Defending against black-box membership inference attacks via adversarial examples, in: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 259–274.
[16] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert and Y. Zhang, When machine unlearning jeopardizes privacy, *arXiv preprint arXiv:2005.02205*, 2020.
[17] L. Song and P. Mittal, Systematic evaluation of privacy risks of machine learning models, *arXiv preprint arXiv:2003.10595*, 2020.
[18] L. Wang, J.P. Near, N. Somani, P. Gao, A. Low, D. Dao and D. Song, Data capsule: A new paradigm for automatic compliance with data privacy regulations, in: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 2019, pp. 3–23.
[19] T. Kraska, M. Stonebraker, M. Brodie, S. Servan-Schreiber and D. Weitzner, Schengendb: A data protection database proposal, in: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 2019, pp. 24–38.
[20] T. Pasquier, D. Eyers and M. Seltzer, From here to provtopia, in: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 2019, pp. 54–67.
[21] J.A. Kroll, N. Kohli and P. Laskowski, Privacy and policy in polystores: A data management research agenda, in: *Heterogeneous Data Management, Poly-stores, and Analytics for Healthcare*, 2019, pp. 68–81.
[22] J. Mohan, M. Wasserman and V. Chidambaram, Analyzing gdpr compliance through the lens of privacy policy, in: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 2019, pp. 82–95.
[23] J. Rogers, J. Bater, X. He, A. Machanavajjhala, M. Suresh and X. Wang, Privacy changes everything, in: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, 2019, pp. 96–111.
[24] M. Kim, J. Lee, L. Ohno-Machado and X. Jiang, Secure and differentially private logistic regression for horizontally distributed data, *IEEE Transactions on Information Forensics and Security* **15** (2019), 695–710.
[25] L. Bonomi, Y. Huang and L. Ohno-Machado, Privacy challenges and research opportunities for genomic data sharing, *Nature Genetics*, 2020, 1–9.
[26] K. Abouelmehdi, A. Beni-Hessane and H. Khaloufi. Big healthcare data: Preserving security and privacy, *Journal of Big Data* **5**(1) (2018), 1.
[27] W.N. Price and I.G. Cohen, Privacy in the age of medical big data, *Nature medicine* **25**(1) (2019), 37–43.
[28] N. Hallowell, M. Parker and C. Nellaker, Big data phenotyping in rare diseases: Some ethical issues, *Genetics in Medicine* **21**(2) (2019), 272–274.
[29] A. Jack et al., *International classification of diseases for oncology: ICD-O*, World Health Organization, 2000.
[30] Y. JKim, Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882*, 2014.
[31] M. Alawad et al., Privacy-preserving deep learning NLP models for cancer registries, *IEEE Transactions on Emerging Topics in Computing*, 2020.
[32] F. Chollet et al., Keras, https://keras.io, 2015.
[33] M. Abadi et al., Tensorflow: A system for large-scale machine learning, in: *12th {USENIX}Symposium on Operating Systems Design and Implementation*, 2016, pp. 7265–283.
[34] J.M. Wozniak et al., Candle/supervisor: A workflow framework for machine learning applied to cancer research, *BMC Bioinformatics* **19**(18) (2018), 59–69.