Extraction of Tumor Site from Cancer Pathology Reports using Deep Filters

Abhishek K Dubey
Biomedical Science, Engineering, and Computing Group,
Oak Ridge National Laboratory
Oak Ridge, TN, USA
dubeyak@ornl.gov

Jacob Hinkle
Biomedical Science, Engineering, and Computing Group,
Oak Ridge National Laboratory
Oak Ridge, TN, USA

J. Blair Christian
Biomedical Science, Engineering, and Computing Group,
Oak Ridge National Laboratory
Oak Ridge, TN, USA

Georgia Tourassi
Biomedical Science, Engineering, and Computing Group,
Oak Ridge National Laboratory
Oak Ridge, TN, USA

ABSTRACT

Purpose: Pathology reports are the primary source of information concerning the millions of cancer cases across the United States. Cancer registries manually process the pathology reports to extract the pertinent information including primary tumor site, behavior, histology, laterality, and grade. Processing a large volume of the pathology reports in a timely manner is a continuing challenge for cancer registries. The purpose of this study is to develop an information extraction pipeline to reliably and efficiently extract reportable information.

Method: We have developed a novel inverse-regression (IR) based information extraction pipeline. The inverse-regression based supervised filter has been successfully applied to many application domains. However, its application to the information extraction from unstructured text is hindered primarily by the extreme high-dimensionality of n-gram representations of text. In this study, we attempt to overcome the obstacles by a novel bootstrapping strategy. First, we use an information-theoretic mutual information based filter to discard the excessive and redundant n-gram features. This step reduces the size and improves the condition number of the sample covariance matrix, thus reducing the computational cost and improving the numerical stability of the subsequent inverse-regression step. Then we use localized sliced inverse-regression (LSIR) to learn a low-dimensional discriminatory subspace for information inference. In particular, we use the k-nearest neighbors of an unlabeled pathology report in the learned representation to infer the desired information from the labeled data in a supervised manner.

Results: The experiments were conducted on a set of de-identified pathology reports with human expert labels as the ground truth. Our pipeline consistently performed better than or comparable to the best performing state-of-the-art methods while reducing the training and inference times substantially.

Conclusion: Our results demonstrate the potential of inverse-regression based information extraction pipeline for reliable and efficient information extraction from unstructured text. The information extracted from the pathology reports can be used along with clinical information, medical imaging, and genomic information to instigate discoveries in cancer research.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342173

CCS CONCEPTS

• **Information systems** → *Information retrieval.*

KEYWORDS

text classification; supervised dimension reduction; minimal redundancy and maximal relevance; localized sliced inverse regression

ACM Reference Format:
Abhishek K Dubey, Jacob Hinkle, J. Blair Christian, and Georgia Tourassi. 2019. Extraction of Tumor Site from Cancer Pathology Reports using Deep Filters. In *Proceedings of the 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Niagara Falls, NY (ACM-BCB '19).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3307339.3342173

1 INTRODUCTION

Large-scale cancer surveillance efforts use pathology reports as a primary source of cancer incidence data in the United States. These reports are collected from pathology laboratories across the USA and maintained by dedicated cancer registries. Cancer registries manually process these reports and extract pertinent information that is essential to monitor cancer incidence trends. The extracted information includes tumor site, histology, laterality, behavior, grade, and metastatic status. Automated extraction of this information from pathology reports is instrumental in achieving cancer surveillance in a timely fashion. In this project, we have developed a novel deep information extraction pipeline.

Information extraction from unstructured text documents is a well-established research field. Early works in this field of study include rule-based pattern matching techniques [3, 9, 13]. The rule-based methods are inadequate for the information extraction task since various text expressions are used in practice to encode the same concept. Besides, the rule-based methods rely on domain experts to identify useful patterns and encode them in the program.

Recent advancements in machine learning have opened new avenues to fully automate the information extraction process. They pose the information extraction task as a supervised learning classification task. The labeled text documents are used to learn a useful document representation and then to train a classifier. After learning, information is inferred from a previously unseen document by embedding it in the learned representation space and then classifying it to one of the possible class labels.

Convolution neural networks (CNN) proposed by [27] for the information extraction task embed the pathology reports into a vector space ($\mathbb{R}^{900}$) and then use a fully-connected feed-forward layer for the classification. The words of the documents are first represented in a vector space ($\mathbb{R}^{300}$) instead of its native high-dimensional one-hot encoding representation, which significantly reduces the number of parameters in the subsequent layers. Convolution filters are then used on the word embeddings to extract the n-gram features from the pathology reports, which are then aggregated by a max-over time pool to find a positional invariant document representation. The network parameters including word embedding, convolutional filters, and feed-forward weights are learned simultaneously through back-propagation. The CNN architecture has gained a lot of attention in text analysis recently and is an active area of research [10, 19, 26, 27]. Some of the outstanding challenges in the area include understanding the document representation, training the network under extreme class imbalance, and developing uncertainty quantification measures.

An Inversion-regression (IR) based approach for the information extraction task has been previously demonstrated by [11]. The pathology reports are first represented by the n-gram features, typically used in natural language processing. Localized sliced inverse-regression [33] is then used for learning a low-dimensional document representation from the labeled data. The learned representation is the projection of the high-dimensional n-gram features onto a linear subspace that enhance the discrimination among the data from different class-labels, similar to discriminant analysis [4]. Section 2.2 includes a detailed review of this method. Then k-nearest neighbors of an unlabeled document in the learned space are used for inferring the label in a supervised way. The inverse-regression based approach has been studied for many other applications [5, 15, 32, 35] and has a strong theoretical basis [8, 21]. The use of the inverse regression for the information extraction task suffers from the high computational cost and numerical instability because of the extreme dimensionality of the n-gram representation of the document. In this study, we use a bootstrapping scheme to overcome both obstacles.

We introduce a mutual information based pre-processing step to preemptively remove the excessive and redundant n-gram features. This step reduces the computational cost of the subsequent inverse-regression step significantly, making it computationally viable with a large number of text documents. Text analysis in general operates in a $p > n$ realm with the n-gram representation. Here p is the cardinality of the document representation and n is the number of available labeled samples. Inverse-regression based methods (similar to many other statistical methods) are numerically unstable when $p >> n$. The numerical instability of the algorithm comes from the ill-posedness of the sample covariance matrix estimates.

Table 1: Summary of label distribution in the de-identified pathology reports

Label type	(Labels, counts)	#documents
Organ	(Lung, 504), (Breast, 521)	1025
ICD-O-3 code	(c34.0, 26), (c34.1, 139), (c34.2, 11), (c34.3, 78), (c34.8, 6), (c34.9, 191), (c50.0, 1), (c50.1, 13), (c50.2, 36), (c50.3, 10), (c50.4, 63), (c50.5, 21), (c50.6, 2), (c50.9, 292)	951
Laterality	(bilateral, 6), (left, 400), (right, 415)	821
Behavior	(metastatic, 9), (in-situ, 101), (malignant, 910), (boderline, 1), (benign, 22)	1043
Grade	(grade 1, 124), (grade 2, 233), (grade 3, 271), (grade 4, 17), (grade 9, 1)	646

By discarding the irrelevant and redundant features in the preprocessing step, we overcome the ill-conditioning of the sample covariance matrix. Our pipeline is shown to produce comparable or better results to state-of-the-art methods at a fraction of the computational cost.

The manuscript is organized as follows: Section 2 describes the dataset, data preprocessing, our method, and evaluation criteria; Section 3 presents the experimental results; Section 4 discusses the potential future directions.

2 MATERIALS AND METHODS

2.1 Material

We use the de-identified pathology reports collected from 5 SEER cancer registries (CT, HI, KY, NM, Seattle). The reports were labeled by state cancer registries, following the standard coding guidelines issued by SEER. The assigned primary-site labels are standard ICD-O-3 topography codes used to encode the tumor site as per the SEER coding guidelines. A summary of the labels including the ICD-O-3 code are provided in Table 1. Further details of the dataset can be found in our previously published work [27].

The pathology reports were provided in XML format. We discarded the meta-data associated with the pathology reports and used only unstructured text sections of the XML file. An example of pathology report (unstructured text section) is included in Figure 1. We used regular expressions to standardize the reports. For instance, we replaced any integer greater than 999 to "largeint", we replaced any floating point number to "floattoken", replaced the "o clock" token as "oclock", and we used white-space and other delimiters such as ".", ":", ";" to tokenize the reports.

2.2 Method

We formulate the extraction of tumor site (ICD-O-3 code) from the pathology reports as a supervised classification task. Supervised classification is one of the widely studied pattern recognition models in machine learning. A classification task relies on the labeled data to build a classifier, which is then used to predict the class-labels for the unlabeled data. In this work, we have developed a novel deep filter pipeline to first learn a low-dimensional document

<clinical info>
age[in 50s]-year-old woman with right breast mass. final dx>place, wa; ***path-number[1] (**date[jan 18 2012])
right breast, 10:00, ultrasound-guided core biopsies (part a): invasive ductal carcinoma with the following features:
1. nottingham grade 3 of 3 derived as follows: poor tubule formation (3), high nuclear grade (3), moderate mitotic activity (2).
2. extent/size: involving three of three tissue cores with a maximum contiguous length of 0.5 cm.
3. associated in situ component: high grade ductal carcinoma in situ with\n central necrosis.
4. microcalcifications: identified associated with invasive carcinoma.
5. angiolymphatic invasion: no definite.
6. prognostic markers were performed at cellnetix and immunostains are available for review:
a. invasive carcinoma is negative for estrogen receptor expression (**name[zzz]\n score 0 of 8, with positive internal controls). dcis has focal, weak er\n staining (**name[zzz] score 3 of 8).
b. invasive carcinoma is negative for progesterone receptor expression (**name[zzz]\n score 0 of 8, with positive internal controls). dcis is negative for pr\n expression (**name[zzz] score 0 of 8).
c. invasive carcinoma is positive for her2/neu over-expression by immunohistochemistry
right breast, 8:00, ultrasound-guided core biopsy (part b): invasive ductal carcinoma with the following features:
1. nottingham grade 3 of 3 derived as follows: poor tubule formation (3),\n high nuclear grade (3), moderate mitotic activity (2).
2. extent/size: involving two of two tissue cores with a maximum contiguous length of 0.4 cm.
3. associated in situ component: not identified.
4. microcalcifications: not identified.
5. angiolymphatic invasion: no definite.
6. prognostic markers were performed at cellnetix and immunostains are available for review:
a. negative for estrogen reception expression (**name[zzz] score 0 of 8, with\n positive internal controls).
b. negative for progesterone receptor expression (**name[zzz] score 0 of 8, with\n positive internal controls).
c. positive for her2/neu over-expression by ihc.\n ms/jps
procedures used to establish the diagnosis:
routine\n per discussion with drs. **name[yyy] and **name[xxx] at bcsc conference on 7/5/2100, if\n additional surgical material is received in the future, repeat hormonal\n studies are requested.amendment reason: this amendment is issued to correct the date of\n specimen receipt. the final diagnosis remains unchanged.amendment reason: this amendment is issued to correct the date of\n specimen receipt. the final diagnosis remains unchanged.materials received:\n label consult accession no blocks/slides description\n a ***path-number[1] 0b,18s right breast mass, right axillary lnmaterials received:\n label consult accession no blocks/slides description\n a ***path-number[1] 0b,18s right breast mass, right axillary ln\n

Figure 1: An exemplary pathology report of a breast cancer patient is shown. The highlighted information are the most relevant texts to extract the organ (breast), ICD-O-3 code (c50.9), laterality (right), behavior (malignant) and grade (3). The figure shows that the most of the texts are excessive and irrelevant for the information extraction task.

representation, which co-locates the documents that belong to the same class-label and separates them otherwise. We then train a k-NN classifier on this representation to infer the class-labels from the unlabeled pathology documents.

The high-dimensionality of data is challenging for traditional classifiers including the k-NN classifier, which require a large volume of labeled data to learn any statistically reliable model. A high-dimensional n-gram features, also popularly known as the *bag of words* representation, are often used to represent text documents. One popular approach to overcome the challenge is to reduce the dimensionality of the representation via unsupervised or supervised dimension reduction methods. The unsupervised methods, including both linear (e.g., PCA, NMF, LPP) and non-linear (e.g., LLE, MDS, ISOMAP) approaches [12], often tend to discard the less prevalent discriminatory features without the supervision from the class-labels. We adopt a supervised approach in this work to overcome the challenge.

Supervised dimension reduction methods are categorized in the literature [7] into the filter, wrapper, and embedding methods. The filter methods rely on the intrinsic properties of representation and its interdependence on the class-labels for the dimension reduction. Some prevalent examples of the filter methods include mRMR [24], FS-score [17], LDA [20], SIR [21], and CCA [29]. In contrast, wrapper methods such as FSV [16] and SVM-RFE [18] use a classifier to score the utility of a given subset of features for the classification task and then selects the subset that maximizes the utility. With a large number of features, the subset search space becomes prohibitively large to navigate in a computationally efficient way. The last category is embedding methods that combine the feature selection and classification task in a single formulation. Elastic-net [36] and Lasso [30] are two popular methods in this category. The accuracy of these embedding methods is seen to deteriorate when the number of features far exceeds the number of labeled samples, as in our case. The introduced method comes under the first category.

Our dimension reduction method is based on the inverse-regression based approach adopted by [11]. A brief review of their method is outlined here. Assume the functional dependence of a response variable $Y \in \mathbb{R}$ on a multivariate explanatory variable $\mathbf{X} \in \mathbb{R}^p$ through an unknown k-dimensional subspace where $k < p$. Given a set of observations $\{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ of $\mathbf{X}$ and Y, the goal of inverse regression is to estimate the basis of the unknown low-dimensional subspace, hereby denoted as $\boldsymbol{\beta} = \{\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_k}\}$, also known as the *central subspace*. The dependence of any observation y_i on $\mathbf{x_i}$ is only through a low-dimensional projection $(\boldsymbol{\beta_1}^T \mathbf{x_i}, \ldots, \boldsymbol{\beta_k}^T \mathbf{x_i})^T$. The key idea of inverse regression is that the

explanatory variable $\mathbf{X}$ are regressed against Y instead of other way round as in regression. As Y varies, the inverse regression curve $\mathbb{E}(\mathbf{X}|Y) - \mathbb{E}(\mathbf{x})$ typically lies in a k-dimensional subspace instead of $\mathbb{R}^p$ under some mild distributional assumption on $\mathbf{X}$. Consequently, the conditional covariance matrix $\mathrm{Cov}(\mathbb{E}(\mathbf{Z}|Y))$ is degenerate in any direction orthogonal to the $\boldsymbol{\beta}$, where $\mathbf{Z}$ is standardized X.

Sliced inverse regression (SIR) [21] was introduced by Li in 1991 to estimate the subspace $\boldsymbol{\beta}$. In general, when the covariance matrix of $\mathbf{X}$ is Σ and the covariance matrix of $\mathbb{E}(\mathbf{X}|Y)$ is $\boldsymbol{\Gamma}$, the unknown *central subspace* $\boldsymbol{\beta}$ can be obtained by solving a generalized-eigenvalue problem,

$$\boldsymbol{\Gamma}\boldsymbol{\beta} = \lambda\Sigma\boldsymbol{\beta}. \tag{1}$$

Given a set of observations $\{(\mathbf{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{n}}$ with the sample mean adjusted to zero (i.e., $\mathbb{E}(\mathbf{X}) = \mathbf{0}$), the SIR [21] divides the range of Y variable into H non-overlapping intervals $\{\mathrm{Y_1}, ..., \mathrm{Y_H}\}$ and groups the observations into H groups $\{\mathrm{G_1}, ..., \mathrm{G_H}\}$ based on their Y values. The sample covariance matrix and conditional covariance matrix are then estimated by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}\,\mathbf{x_i}^T, \tag{2}$$

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{n}\sum_{h=1}^{H} |G_h| \left(\frac{1}{|G_h|}\sum_{\mathbf{x_j}\in G_h} \mathbf{x_j}\right)\left(\frac{1}{|G_h|}\sum_{\mathbf{x_j}\in G_h} \mathbf{x_j}\right)^T, \tag{3}$$

where $G_h = \{\mathbf{x_i} : \mathbf{x_i}$ belongs to the h-th slice of the range of the Y variable$\}$ and $|G_h|$ is the size of set G_h. When $p > n$ or the explanatory variables are highly collinear then the sample covariance matrix $\hat{\Sigma}$ is singular. The precursor work [11] uses a ridge regularization to overcome the rank-deficiency by adding a diagonal matrix $s\mathbf{I}_p$ to the sample covariance matrix in (1), where s is the ridge regularization parameter and $\mathbf{I}_p$ is the identity matrix. Also the precursor work [11] uses a variant of the regularized SIR method, in particular they used the localized sliced inverse regression (LSIR) [33]. The SIR assumes the elliptical distribution of the explanatory variables $\mathbf{X}$ for a given Y, which is relaxed in the LSIR to support the multimodal elliptical distribution. In many practical scenarios, the global-slice mean $\mathbb{E}(\mathbf{X}|Y_h))$ is bad characterization of the inverse regression curve. The LSIR uses the local-slice mean estimate to characterize the inverse regression curve. The conditional covariance matrix $\hat{\boldsymbol{\Gamma}}$ is thus computed with the local-slice mean estimate by

$$\hat{\boldsymbol{\Gamma}}_{loc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{m_{i,loc}}\,\mathbf{m_{i,loc}}^T, \tag{4}$$

where $\mathbf{m_{i,loc}}$ is the local-slice mean estimate at $\mathbf{x_i}$ computed over k-nearest neighbors

$$\mathbf{m_{i,loc}} = \frac{1}{|\mathcal{N}(\mathbf{x}_i)|}\sum_{\mathrm{j}\in\mathcal{N}(\mathbf{x_i})} \mathbf{x_j}, \tag{5}$$

where $\mathcal{N}(\mathbf{x_i}) = \{\mathbf{x_j} : \mathbf{x_j}$ belongs to the k-nearest neighbors of $\mathbf{x_i}$ in G_h, where G_h is the slice to which $\mathbf{x_i}$ belongs$\}$. A complete outline of the dimension reduction procedure with the LSIR is as follows.

(1) Compute the eigenvalue decomposition of $(\hat{\Sigma}+s\mathbf{I}_p)^{-1}\hat{\boldsymbol{\Gamma}}_{loc} = \mathbf{P}\Lambda\mathbf{P}^{-1}$, where $\mathbf{P}$ and Λ are the eigenvectors and eigenvalues respectively.
(2) Sort the eigenvalues Λ in the descending order and permute the eigenvectors accordingly in descending order. Let Λ_s and $\mathbf{P}_s$ denote the sorted eigenvalues and permuted eigenvectors.
(3) Project the dataset onto the leading r dimensional subspace and scale by the eigenvalues as $\mathbf{x}_r = \Lambda_r^{\frac{1}{2}}\mathbf{P}_r^T\mathbf{x}$, where $\mathbf{P}_s = [\mathbf{P}_r \;\; \mathbf{P}_{p-r}]$ and $\Lambda_s = [\Lambda_r \;\; \Lambda_{p-r}]$.

Building upon the precursor work [11], we have introduced an inverse-regression based pipeline for the information extraction task, which is composed of two filter stages. At the first stage, we use a computationally efficient minimal redundancy and maximal relevance (mRMR) [24] filter to select the top k features that are relevant for the discrimination task and have minimum redundancy. At the second stage, we use the localized sliced inverse regression (LSIR) [33] for subspace learning to find a low-dimensional document embedding. By introducing the mRMR filter prior to the inverse-regression step, the excessive and redundant features are discarded in advance, which substantially reduce the computational cost of the subsequent inverse-regression step. The estimation of $(\hat{\Sigma}+s\mathbf{I}_p)^{-1}$ and eigendecomposition of $(\hat{\Sigma}+s\mathbf{I}_p)^{-1}\hat{\boldsymbol{\Gamma}}_{loc}$ are the two most computationally expensive steps of the inverse regression with computational complexity $O(p^3)$ in practice for direct methods [31]. With the mRMR pre-processing, the size of $(\hat{\Sigma}+s\mathbf{I}_p)$ and $(\hat{\Sigma}+s\mathbf{I}_p)^{-1}\hat{\boldsymbol{\Gamma}}_{loc}$ is reduced from $p\times p$ to $k\times k$, thus reducing the computational cost of these steps from $O(p^3)$ to $O(k^3)$. The first step also improves the condition number of the sample covariance matrix $\hat{\Sigma}$ by increasing the sample-to-feature ratio (n/p), thus improving the numerical stability of the subsequent step. Next, we describe the introduced pipeline outlined below in detail except the inverse-regression step.

(1) Document representation by the n-gram features
(2) Minimal redundancy and maximal relevance filter for removing the excessive and redundant features
(3) Localized sliced inverse regression for learning a low-dimensional discriminatory subspace
(4) k-nearest neighbor classifier on a low-dimensional subspace to infer the class-labels from the unlabeled data

2.2.1 Document representation. We represent the cancer pathology reports in vector space by the n-gram features, which capture short sequence information and ignores longer ones. In particular, we use the log-normalized count of each short word sequence t, $\log(1 + tf(t, d))$, as a feature, where $tf(t, d)$ denotes the count of word sequence t in the document d. The n-gram representation has proven useful for many natural language processing (NLP) tasks in the literature [14]. Term-frequency and inverse-document frequency (tf-idf) is another popular choice for document representation [28]. Traditional approaches retain only the top n-gram features to cope with the high-dimensionality of the document representation for post-analysis, which is ill-suited for the information extraction task because the less prevalent but relevant features that contain the information to be extracted are discarded. In this work, we use a supervised dimension reduction pipeline as an alternative.

2.2.2 Minimal redundancy and maximal relevance filter. We first use mutual information based information-theoretic filter to discard excessive and redundant features. Mutual information measures the interdependence between two random variables quantifying the amount of information obtained about one variable observing the other. Given the feature random variables $\mathbf{X} = (X_1, ..., X_p)^T$ and class-label random variable Y, a subset of the feature variables, $\mathbf{X}_s \subset \mathbf{X}$, of maximal cardinality k that holds the maximal information regarding the class-label, can be obtained, in theory, by solving the combinatorial problem,

$$\hat{\mathbf{X}}_s = \arg\max_{\mathbf{X}_s} \mathbf{I}(\mathbf{X}_s; Y), \quad s.t. \ \ \mathbf{card}(\mathbf{X}_s) \leq k, \tag{6}$$

where $\mathbf{I}(\mathbf{X}_s; Y)$ denotes the joint mutual information of the selected feature variables $\mathbf{X}_s$ and the class-label variable Y, defined as,

$$\mathbf{I}(\mathbf{X}; Y) = \int_{\mathbf{X}} \int_{\mathbf{Y}} p(\mathbf{x}, y) \log \left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(y)} \right) \mathrm{dyd}\mathbf{x}, \tag{7}$$

where $p(\mathbf{x}, y)$ is the joint multi-variate density estimate of $(\mathbf{X}_s, Y)$, and $p(\mathbf{x})$ and $p(y)$ are the marginal density estimates of $\mathbf{X}_s$ and Y respectively. Obtaining the exact solution of (6) is difficult in practice for two reasons. First, the problem (6) is known to be NP-hard – i.e., there is no known efficient (tractable) algorithm to solve large instances of (6) to date [24]. Second, the estimation of $p(\mathbf{x})$ and $p(\mathbf{x}, y)$ is known to be ill-posed in the high-dimensional setting as it often requires the inversion of the sample covariance matrices, which are rank-deficient when the number of features are more than the number of labeled samples.

Minimal redundancy and maximal relevance (mRMR) method, introduced by Peng et al. [24], provides an efficient greedy solution to the problem formulation (6). Let $\mathbf{S} = \{X_{S_1}, \ldots, X_{S_{t-1}}\}$ be the set of selected features at time step $t-1$. The mRMR method greedily selects the next feature by solving the following optimization problem,

$$X_{S_t} = \arg\max_{i \notin \mathbf{S}} \mathbf{I}(X_i; Y_c) - \frac{1}{t-1} \sum_{k=1}^{t-1} \mathbf{I}(X_{S_k}; X_i). \tag{8}$$

The mRMR method is shown by the authors to be optimal for the online version of the problem (6). In the online setting, $t-1$ features are given and the t-th feature is selected to maximize the joint mutual information of (6). The first term of the mRMR method (8) is the mutual information gain with the new selection X_i, whereas the second term is the average redundancy in the mutual information with the previous selections. The mRMR method maximizes the mutual information gain through including the new feature in the selected cohort while minimizing the average redundancy with the already selected cohort $\mathbf{X}_{S^{t-1}}$. We use the mRMR filter to select the top k features that are relevant for the discrimination task and have minimum redundancy.

2.2.3 Localized sliced inverse regression. Described earlier in Section 2.2.

2.2.4 Classifier. We use the document embedding obtained from the LSIR to train a k-nearest neighbor classifier for tumor site classification, which is then used to infer the tumor site from the unlabeled data. The number of neighbors parameter (k) for the inference is set empirically.

2.3 Experimental setup and evaluation metric

We compare the classification accuracy of the inverse regression based methods with the state-of-the-art deep learning based method [27]. We use tenfold cross-validation to partition the dataset into 10 parts with nearly balanced label distributions among the partitions. We use 9 partitions for training the classifier and use 1 partition for the model evaluation. We report micro F-score, along with macro- average F-score, precision and recall.

$$F_{\text{micro}} = \frac{P_{\text{micro}} R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}, \tag{9}$$

$$P_{\text{micro}} = \frac{\sum_{C_j} TP_j}{\sum_{C_j} TP_j + FP_j}, \quad R_{\text{micro}} = \frac{\sum_{C_j} TP_j}{\sum_{C_j} TP_j + FN_j}$$

$$P_{\text{macro}} = \frac{1}{|C|} \cdot \sum_{C_j} P(C_j), \quad P(C_j) = \frac{TP_j}{TP_j + FP_j} \tag{10}$$

$$R_{\text{macro}} = \frac{1}{|C|} \cdot \sum_{C_j} R(C_j), \quad R(C_j) = \frac{TP_j}{TP_j + FN_j} \tag{11}$$

$$F_{\text{macro}} = \frac{1}{|C|} \cdot \sum_{C_j} \frac{2P(C_j)R(C_j)}{P(C_j) + R(C_j)} \tag{12}$$

Here TP_j denotes the number of true positives for class C_j, FP_j denotes the number of false positives for class C_j, and FN_j denotes the number of false negatives for class C_j. Micro-average measures show how well the classifier performs overall, whereas macro-average measures how well the classifier performs across all classes, irrespective of class prevalence. We also report the confusion matrix for detailed examination of misclassifications.

3 RESULTS

We compare the classification results of our method (Section 2) for the tumor site classification task against both traditional machine

Figure 2: Word cloud chart of the top features selected by the mRMR method. Larger tokens in the word cloud chart are ranked higher by the mRMR method.

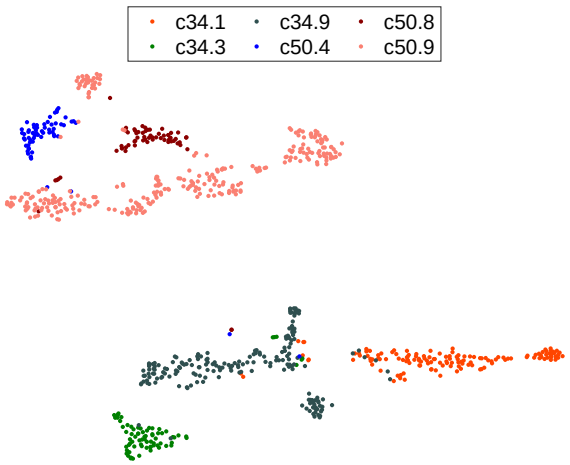


Figure 3: Visualization of 825 de-identified pathology reports in the central subspace with the t-SNE [23]. The top 2000 selected features of the pathology reports by the mRMR method are projected into a 12-dimensional subspace by the LSIR with the regularization parameter s set to $5/n$, where n is the number of pathology reports. The pathology reports with the same ICD-O-3 topographical code are colored in the same color.

learning algorithms and a precursor neural network method [27]. Among traditional methods, we compare against Naive Bayes (NB), logistic regression (LR), and support vector machine (SVM). We use the top 400 unigram and bigram term frequency-inverse document frequency (tf-idf) features for these methods, as in the precursor work [27], which gives the best results among the other examined settings. We compare also against the state-of-the-art convolutional neural network (CNN) method [27]. In particular, we compared with two CNN models: (i) a CNN+PubMed model where a pre-trained (fixed) PubMed word embedding is used while learning the other network parameters (convolutional filters and feedforward weights), (ii) a CNN+None model where the word embeddings are jointly learned with the convolutional filters and feedforward classifier weights.

We compare our method, hereby referred to mRMR+LSIR+kNN, also against the precursor method [11], referred to LSIR+kNN. For both methods, the n-gram features (see Section 2.2.1) up to length 4 including the unigram, bigram, trigram, and quadgram features are used to represent the reports. Figure 2 shows the top n-gram word sequences selected by the mRMR method. Figure 3 shows the 2D t-SNE embedding of the pathology reports in the central subspace learned with the LSIR method following the mRMR feature selection step. The pathology reports with the same class-labels are seen to cluster together in this low-dimensional document representation.

In the experiments with the mRMR+LSIR+kNN method, we remove the n-gram features that occur in less than 4 reports, which gives a 28946-dimensional representation of the de-identified reports. We first use the mRMR method to select top 2000 features, and then use the LSIR with the regularization parameter set to $5/n$ to learn a 12-dimensional representation. The number of neighbors

Table 2: Micro- and macro- average scores for the tumor site classification task on the test pathology reports obtained by tenfold cross-validation partition of a corpus of 825 de-identified reports that consists of 6 prevalent classes (each class containing at least 50 samples). The results of other methods except for the mRMR+LSIR+kNN are taken from the precursor study [11] for comparison.

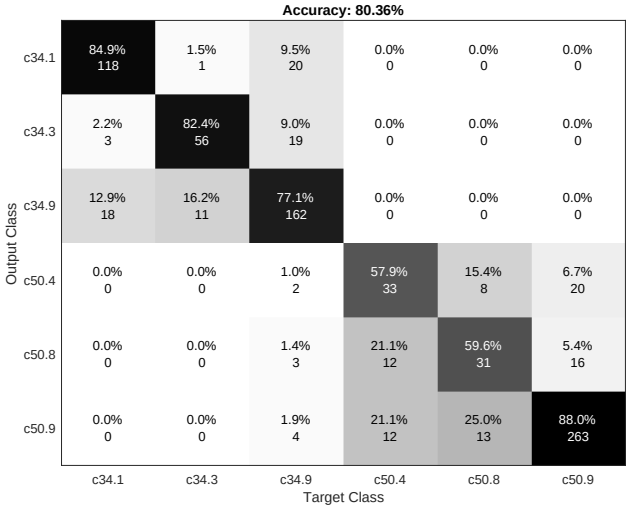
Approach	F_{micro}	F_{macro}	P_{macro}	R_{macro}
NB	0.679	0.475	0.500	0.500
LR	0.744	0.546	0.587	0.552
SVM	0.760	0.640	0.684	0.625
CNN+PubMed	0.797	0.688	0.733	0.695
CNN+None	0.811	0.701	0.737	0.707
LSIR+kNN	0.804	0.734	0.769	**0.787**
mRMR+LSIR+kNN	**0.821**	**0.759**	**0.775**	0.746

for the conditional covariance matrix estimation (see Equation (4)) in the LSIR method is set to 20, and the number of neighbors in the k-nearest neighbor classifier is set to 5. All these parameters are empirically set. In the LSIR+kNN method, we use the same initial representation and LSIR parameter values as used in the precursor work [11], which gives the best results among the other explored settings.

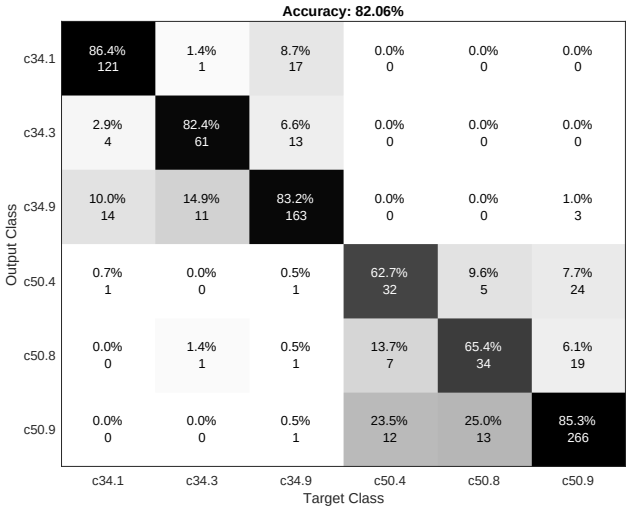
Table 2 shows a comparison of the classification accuracy of the inverse-regression based methods with the other methods for the tumor site classification task. We report the classification results with a corpus of 825 de-identified reports that consists of 6 prevalent classes, each of which has at least 50 samples. The results show that the inverse-regression based methods achieve higher macro-average and micro- average scores than the other methods. The MRMR+LSIR+kNN method also performs better than the LSIR+kNN method while also reducing the computation time by more than 10x as observed on a MacBook Pro 2018 with 2.7 GHz Intel Core i7 processor. Figure 4 shows the confusion matrix for the inverse-regression based methods.

4 DISCUSSION

We have developed a supervised deep pipeline for embedding unstructured pathology reports to a low-dimensional vector space for information extraction of the primary cancer site. This pipeline is built upon the precursor work [11] that demonstrated the effectiveness of inverse regression for the information extraction. The two most computationally expensive steps in the inverse regression are computing the inverse of $(\hat{\Sigma} + s\mathbf{I}_p)$ and finding the eigenvectors of $(\hat{\Sigma} + s\mathbf{I}_p)^{-1}\hat{\Gamma}_{loc}$, both of which cost $O(p^3)$ with direct methods. With millions of pathology reports and n-grams of length up to 4, the number of n-gram features (p) could grow to tens of millions, making the proposed method prohibitively expensive. In this work, we have demonstrated the use of the mRMR method as a pre-processing step to discard the irrelevant and redundant features upfront, thus reducing the computational cost to $O(k^3)$, where k is the selected features. We showed that the introduction of this pre-processing step also improves the overall accuracy of the task as the pre-processing step improves the condition number of the



(a) LSIR+kNN



(b) mRMR+LSIR+kNN

Figure 4: Confusion matrix for the tumor site classification with the inverse-regression based methods on the test pathology reports obtained by tenfold cross-validation partition of 825 **de-identified data that has at least** 50 **samples per class-label.**

sample covariance matrix by improving the sample-to-feature ratio from n/p to n/k.

Moving forward, we would like to address other open challenges in this field of study. Dispersion of the features because of the frequent misspellings and liberal (diverse) usages of language to express the same concept is a continuing challenge. Many relevant features become weakly correlated with the class-labels due to the dispersion. In future work, we would like to exploit the local contexts and topic structure of the n-grams to identify the dispersion and would like to develop a supervised information-theoretic technique to efficiently aggregate the dispersed features in a preprocessing step to the supervise dimension reduction step. The local context and topic structure has been successfully exploited previously in text analysis in various ways [1, 2, 6, 22, 25, 34]. Bringing awareness of these structures to information extraction would be a valuable addition to this pipeline, which uses the interdependence between the n-gram features and class-labels to obtain a low-dimensional discriminatory document representation.

5 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.

The authors wish to thank Valentina Petkov of the Surveillance Research Program from the National Cancer Institute and the SEER registries for providing the de-identified pathology reports. The authors wish to thank Folami T. Alamudun, and Shang Gao at ORNL for sharing their expertise in the field of study.

REFERENCES

[1] Adnan Ahmad and Mohammad Ruhul Amin. 2016. Bengali word embeddings and it's application in solving document classification problem. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 425–430.

[2] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, 1 (2015).

[3] Douglas E Appelt and Boyan Onyshkevych. 1998. The common pattern specification language. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*. Association for Computational Linguistics, 23–30.

[4] Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* 18 (1998), 1–8.

[5] Caroline Bernard-Michel, Sylvain Douté, Mathieu Fauvel, Laurent Gardes, and Stephane Girard. 2009. Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research: Planets* 114, E6 (2009).

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[7] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.

[8] R Dennis Cook and Liqiang Ni. 2005. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* 100, 470 (2005), 410–428.

[9] Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities* 36, 2 (2002), 223–254.

[10] Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 69–78.

[11] Abhishek Dubey, Hong-Jun Yoon, and Georgia Tourassi. in press. Inverse Regression for Extraction of Tumor Site from Cancer Pathology Reports. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE.

[12] Daniel Engel, Lars Hüttenberger, and Bernd Hamann. 2012. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[13] David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10, 3-4 (2004), 327–348.

[14] Johannes Fürnkranz. 1998. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence* 3, 1998 (1998), 1–10.

[15] Ali Gannoun, Stéphane Girard, Christiane Guinot, and Jerôme Saracco. 2004. Sliced inverse regression in reference curves estimation. *Computational statistics*

& data analysis 46, 1 (2004), 103–122.

[16] Guillermo L Grinblat, Javier Izetta, and Pablo M Granitto. 2010. Svm based feature selection: Why are we using the dual?. In *Ibero-American Conference on Artificial Intelligence*. Springer, 413–422.

[17] Quanquan Gu, Zhenhui Li, and Jiawei Han. 2012. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725* (2012).

[18] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1-3 (2002), 389–422.

[19] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. 2018. Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access* 6 (2018), 23253–23260.

[20] Peter A Lachenbruch and M Goldstein. 1979. Discriminant analysis. *Biometrics* (1979), 69–85.

[21] Ker-Chau Li. 1991. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* 86, 414 (1991), 316–327.

[22] Qiang Ma and Katsumi Tanaka. 2005. Topic-structure-based complementary information retrieval and its application. *ACM Transactions on Asian Language Information Processing (TALIP)* 4, 4 (2005), 475–503.

[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[24] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 8 (2005), 1226–1238.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems* 108 (2016), 42–49.

[27] John X Qiu, Hong-Jun Yoon, Paul A Fearn, and Georgia D Tourassi. 2018. Deep learning for automated extraction of primary sites from cancer pathology reports. *IEEE journal of biomedical and health informatics* 22, 1 (2018), 244–251.

[28] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.

[29] Bruce Thompson. 1984. *Canonical correlation analysis: Uses and interpretation*. Number 47. Sage.

[30] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[31] Lloyd N Trefethen and David Bau III. 1997. *Numerical linear algebra*. Vol. 50. Siam.

[32] YH Tu, ZN Fu, Ao Tan, Gan Huang, L Hu, YS Hung, and ZG Zhang. 2018. A novel and effective fMRI decoding approach based on sliced inverse regression and its application to pain prediction. *Neurocomputing* 273 (2018), 373–384.

[33] Qiang Wu, Feng Liang, and Sayan Mukherjee. 2010. Localized Sliced Inverse Regression. *Journal of Computational and Graphical Statistics* 19, 4 (2010), 843–860.

[34] Qingqiang Wu, Caidong Zhang, Qingqi Hong, and Liyan Chen. 2014. Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science* 40, 5 (2014), 611–620.

[35] Ting Zhang, Wenhua Ye, and Yicai Shan. 2016. Application of sliced inverse regression with fuzzy clustering for thermal error modeling of CNC machine tool. *The International Journal of Advanced Manufacturing Technology* 85, 9-12 (2016), 2761–2771.

[36] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* 67, 2 (2005), 301–320.