Journal of Biomedical Informatics 122 (2021) 103872

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Supervised line attention for tumor attribute classification from pathology reports: Higher performance with less data

Nicholas Altieri[a,*,1], Briton Park[a,1], Mara Olson[b], John DeNero[c], Anobel Y. Odisho[d,e], Bin Yu[a,c,f,*,1]

[a] Department of Statistics, University of California, Berkeley, United States
[b] UCSF School of Medicine, United States
[c] Department of Electrical Engineering and Computer Science, University of California, Berkeley, United States
[d] Department of Urology, University of California, San Francisco Helen Diller Family Comprehensive Cancer Center, United States
[e] Center for Digital Health Innovation, University of California, San Francisco, United States
[f] Chan-Zuckerberg Biohub, San Francisco, CA, United States

ARTICLE INFO

Keyword:
Natural Language Processing
Cancer
Information Extraction
Pathology
EHR

ABSTRACT

Objective: We aim to build an accurate machine learning-based system for classifying tumor attributes from cancer pathology reports in the presence of a small amount of annotated data, motivated by the expensive and time-consuming nature of pathology report annotation. An enriched labeling scheme that includes the location of relevant information along with the final label is used along with a corresponding hierarchical method for classifying reports that leverages these enriched annotations.

Materials and methods: Our data consists of 250 colon cancer and 250 kidney cancer pathology reports from 2002 to 2019 at the University of California, San Francisco. For each report, we classify attributes such as procedure performed, tumor grade, and tumor site. For each attribute and document, an annotator trained by an oncologist labeled both the value of that attribute as well as the specific lines in the document that indicated the value. We develop a model that uses these enriched annotations that first predicts the relevant lines of the document, then predicts the final value given the predicted lines. We compare our model to multiple state-of-the-art methods for classifying tumor attributes from pathology reports.

Results: Our results show that across colon and kidney cancers and varying training set sizes, our hierarchical method consistently outperforms state-of-the-art methods. Furthermore, performance comparable to these methods can be achieved with approximately half the amount of labeled data.

Conclusion: Document annotations that are enriched with location information are shown to greatly increase the sample efficiency of machine learning methods for classifying attributes of pathology reports.

1. Objective

By enabling patients to receive tailored risk assessment and treatment decisions, precision medicine has the potential to improve healthcare quality. [1] However, effective delivery of precision medicine depends on accurate and detailed patient data. Unfortunately, much of the relevant clinical data, such as cancer stage and histology, are stored as free text in lengthy unstructured or semi-structured reports. [2] Leveraging the data contained in these reports for precision medicine applications relies on manual efforts by annotators with domain expertise for many downstream automated methods. Due to the time-consuming and expensive nature of manual information extraction, researchers and clinicians have worked to develop algorithms to automatically extract pertinent data from pathology reports with mixed success, with machine learning-based methods underlying some of the more effective solutions [2,3]. However, generating sufficient training data for different cancer types is challenging, due to the large number of data elements and their specificity, as well as the need for highly trained annotators. This is a substantial obstacle for automatically structuring biomedical text across clinical conditions and healthcare facilities. Thus, it is critical to develop methods that can provide high accuracy using small training sets.

[*] Corresponding authors at: 367 Evans Hall, #3860, Berkeley, CA, United States.
E-mail addresses: naltieri@berkeley.edu (N. Altieri), binyu@berkeley.edu (B. Yu).

[1] Equal contribution.

https://doi.org/10.1016/j.jbi.2021.103872
Received 19 December 2020; Received in revised form 8 March 2021; Accepted 18 July 2021
Available online 16 August 2021

In this work, we develop a novel hierarchical annotation and corresponding classification method to address the need for high accuracy methods in the presence of a small amount of annotated data. We apply this method to classifying tumor attributes from 250 colon cancer pathology reports and 250 kidney cancer reports at the University of California, San Francisco. Compared to state of the art approaches, we find that our methods typically require half the labeled data to achieve the same level of performance.

2. Background

The abundance of textual data in the clinical domain has led to increased interest in developing biomedical information extraction systems. These systems aim to automatically extract pre-specified data elements from medical documents, such as physician notes, radiology reports, and pathology reports, and store them in databases. Converting the originally free-text data into a structured form makes them easily available to clinical practitioners or researchers.

For categorical attributes, the information extraction task can be viewed as a form of document classification that classifies the value based on document contents. For a given attribute, the value is one of a fixed set of options selected based on information in the document. As an illustration, the set of values for the attribute "presence of lymphovascular invasion" could consist of the values "present" and "absent". Both classical and deep learning classification methods have been applied to this task in the prior work discussed below.

There has been success in applying classical machine learning techniques to classifying attributes of tumors from pathology reports. Yala et al. classified over 20 binary attributes from breast cancer pathology reports using boosting over n-gram features. [3] Jouhet et al. investigated applications of Support Vector Machines (SVMs) and Naive Bayes classifiers to the task of predicting International Classification of Diseases for Oncology (ICD-O-3) from cancer pathology reports. [4]. More recently, there has been success in applying deep learning techniques to pathology report classification. Qiu et al. applied convolutional neural networks (CNNs) to predicting ICD-O-3 from breast and lung cancer pathology reports. [5] Gao et al. applied hierarchical attention networks to predict tumor site and grade from pathology reports within the NCI-SEER dataset and noted improvement in micro-f1 of up to 0.2 compared to baselines across primary site and histologic grade for lung cancer and breast cancer reports.

There has also been work addressing pathology report classification in the absence of a large amount of labeled data. Odisho et al. analyzed performance of machine learning methods for extracting clinical information from prostate pathology reports across various data regimes and found that, while deep learning performed best when trained on the full dataset of 2,066 labeled documents and achieved a mean weighted-F1 score of 0.97 across classification attributes, simpler methods such as logistic regression and adaBoost performed best in smaller data regimes (<256 reports). [6] Additionally, Zhang et al, investigated the problem of unsupervised adaptation across attributes in breast cancer pathology reports. [7] Given a set of attributes with labels and a new attribute without labels but with relevant keywords, they used adversarial adaptation with semi-supervised attention to extract data. We use all of the above methods as baselines for our system to compare against, with the exception of Zhang et al. due to the difference in tasks.

3. Materials and methods

3.1. Data sources

Our data consists of 250 colon cancer pathology reports and 250 kidney cancer reports from 2002 to 2019 at the University of California, San Francisco. The data was split into two sets, a set of 186, which we used for training and validation, and a test set of size 64. We list the tumor attributes and their corresponding possible values in Table 1. Institutional Review Board approval was obtained for this study.

Table 1
Extracted attributes and their possible values.

Attribute	
Tumor Site	
Colon	Cannot be determined, cecum, colon not otherwise specified, hepatic flexure, ileocecal valve, left descending colon, other, rectosigmoid junction, rectum, right ascending colon, sigmoid colon, splenic flexure, transverse colon, or not reported
Kidney	Upper pole, middle pole, lower pole, other, not specified, or not reported
Histologic Type	
Colon	Adenocarcinoma, adenosquamous carcinoma, carcinoma, type cannot be determined, large cell neuroendocrine carcinoma, medullary carcinoma, micropapillary carcinoma, mucinous adenocarcinoma, neuroendocrine carcinoma poorly differentiated, other histologic type not listed, serrated adenocarcinoma, signet-ring cell carcinoma, small cell neuroendocrine carcinoma, squamous cell carcinoma, undifferentiated carcinoma, or not reported
Kidney	Acquired cystic disease associated renal cell carcinoma, chromophobe renal cell carcinoma, clear cell papillary renal cell carcinoma, clear cell renal cell carcinoma, collecting duct carcinoma, hereditary leiomyomatosis and renal cell carcinoma-associated renal cell carcinoma, mit family translocation renal cell carcinoma, mucinous tubular and spindle renal cell carcinoma, multilocular cystic clear cell renal cell neoplasm of low malignant potential, oncocytoma, other histologic type, papillary renal cell carcinoma, papillary renal cell carcinoma type 1, papillary renal cell carcinoma type 2, renal cell carcinoma unclassified, renal medullary carcinoma, succinate dehydrogenase sdh deficient renal cell carcinoma, t611 renal cell carcinoma, tubulocystic renal cell carcinoma, xp11 translocation renal cell carcinoma, or not reported
Procedure	
Colon	Abdominoperineal resection, left hemicolectomy, low anterior resection, not specified, other, polypectomy, right hemicolectomy, sigmoidectomy, total abdominal colectomy, transanal disk excision, transverse colectomy, or not reported
Kidney	Total nephrectomy, partial nephrectomy, radical nephrectomy, other, or not reported
Laterality	
Colon	Not applicable to colon cancer
Kidney	Left, right, or not reported
Grade	
Kidney, Colon	Grade 1, 2, 3, 4, not applicable, or not reported
Lymphovascular Invasion	
Kidney, Colon	Present, absent, or not reported
Perineural Invasion	
Colon	Present, absent, or not reported
Kidney	Not applicable for kidney cancer

3.2. Data annotation methods

Pathology reports consist of free text describing a patient's clinical history and attributes describing the excised specimen, such as surgical procedure, cancer stage, tumor histology, grade, cell differentiation, and presence of invasion to surrounding tissues. More recent pathology reports also contain a synoptic comment section, which is a condensed semi-structured summary of relevant cancer attributes. While many of the most clinically important attributes are reported in this synoptic comment, but not always. All attributes in the College of American Pathology reporting guidelines are annotated for each cancer [9], but for this paper we restrict our investigation to attributes for which some label appears in at least 90% of reports. We annotate our documents using the Multi-document Annotation Environment. [8]

3.3. Enriched annotations

In previous work, annotations consisted of only the label for each attribute in a document. [3,6,10] However, in this work, for each attribute of interest the annotator highlighted all occurrences relevant to the label throughout the document, in addition to the label itself. This provides us with the specific location within the text that directly indicates the attribute's label. Each highlight is classified into the corresponding College of American Pathologists (CAP)-derived category. We investigate two types of annotation: the first we refer to as the "reduced annotation set", a minimal set of annotations containing the line of a given attribute value's first occurrence in the synoptic comment, or, if not in the synoptic comment, the line of where that information is referenced elsewhere in the document. The incremental time required to annotate this location is marginal because the annotator does not need to read any more of the document than that required to annotate the first occurrence of the attribute value. In fact, "We investigated the amount of additional time required to create these enriched annotations and found that it took 20 percent longer on average, primarily due to the time it took the annotator to navigate the attribute drop-down menu. This could perhaps be improved through user interface (UI) considerations. In addition to a reduced annotation set, we also investigate performance with all the occurrences relevant to the final classification highlighted, a more laborious annotation scheme. For our results, unless stated otherwise, we are using the reduced annotation set due to its comparable annotation time to labeling the attribute values alone.

3.4. Data preprocessing

For all methods, we replace all words that occur fewer than two times in the training data with a special $<$ UNK $>$ token, and remove commas, backslashes, semi-colons, tildes, periods, and the word "null" from each report in the corpus. For colons, forward slashes, parentheses, plus, and equal signs, we added a space before and after the character. The spaces were artificially added to preserve semantic value important to the task. For instance, colons often appear in the synoptic comment, and so if an n-gram contains a colon, it can indicate that the n-gram contains important information. If multiple labels for an attribute occurred within a report, we concatenate them to form a single super label. For example, if the report contains both grade 1 and grade 2 as labels for histologic grade, we label the histologic grade of the report as "grade 1 and grade 2".

3.5. Baselines

For all classical baselines, we represent each document as a union of a set of n-grams where n varies from 1 to N, where N is a hyperparameter. For all methods we use random search [11] with 40 trials to tune our hyperparameters according to the 4-fold cross validation error which we found in preliminary experiments to be a good compromise between performance and computational efficiency.

Logistic regression

We use sklearn's [12] logistic regression model with L1 regularization and the liblinear solver. We use balanced class weights to up-weight the penalty on rare classes. We generate 500 points from -6 to 6 logspace for the regularization penalty, and sample 40 points at random.

Support vector classifier

We use sklearn's SVC model with balanced class weights. We define our parameter space as 500 points evenly generated from -6 to 6 in log space for the error penalty C of the model; the kernel as linear or rbf; and the parameter of the kernel as either 0.001, 0.01, 0.1, or 1. We then sample 40 points at random from this space.

Boosting

We use sklearn's adaboost classifier with decision trees of depth 3 and with the SAMME.R boosting algorithm. Our parameter space is 500 points generated evenly from -4 to 1 in logspace for the learning rate and either 25, 50, 100, 25, or 500 for the number of estimators. We then sample 40 points at random from this space.

Hierarchical Attention Network

We implement the hierarchical attention method from Gao et al. This model represents the document as a series of word-vectors. For each sentence in the document it runs a gated recurrent unit (GRU) [13] over the word vectors. It then uses an attention module to create a sentence representation as a sum of the attention-weighted outputs of the GRU. To generate the document representation, a GRU is run over the sentence representations, followed by another attention module is applied to the GRU outputs. The document representation is the attention-weighted sum of the GRU outputs.

For our hyperparameters we use random search across the learning rate, which is either 1e-2, 1e-3, or 1e-4; the width of the hidden layer of the attention module, which is either 50, 100, 150, 200, 250, or 500; the hidden size of the GRU, which is either 50, 100, 150, 200, 250, or 500; and the dropout rate applied to the document representation, which is either 0, 0.2, 0.4, 0.6, or 0.8. We use a batch size of 64 and ADAM [14] as our optimizer.

3.6. ***Our*** *method:* ***Supervised line attention***

In order to take advantage of annotations enriched with location data, we propose a two-stage prediction procedure in which we first predict which lines in the document contain relevant information. We then concatenate the predicted relevant lines and use this string to make the final class prediction using logistic regression.

Finding relevant lines

The first stage predicts which lines are relevant to the attribute. We do this by training an xgboost [15] binary classification model that takes a line represented as a bag of n-grams as its input and outputs whether or not the line is relevant to the attribute. The relevance of each line is predicted independently by this initial classifier.

We then take the top-k lines with the highest scores under the model (where k is a hyperparameter). Groups of adjacent lines are combined into one line so that sentences which span multiple lines are presented to the model as a single line.

Finally, we represent each line as a set of n-grams vectors and compose a document representation as the weighted sum of each vector representation, which is weighted by the score of that line under the xgboost model. If a line is conjoined, its weight is the maximum of all the xgboost scores for each line in the conjoined line. Mathematically, this is represented as

$$d_r(l_1, ..., l_n) = \sum\nolimits_{l_i \in S_k} v(l_i) m(l_i)$$

where $d_r$represents the vector representation of a document d,S_k are the top-k lines with the highest scores under the xgboost model, v is the mapping from a line $l_i$to its set of n-grams representation, and $m(l_i)$ is the xgboost score for line l_i.

With this final weighted representation, we train an L1 regularized logistic regression model with balanced class weights to predict the final class.

We refer to this method as "supervised line attention" due to its relationship to supervised attention in the deep learning literature which predicts relevant locations and creates a weighted representation of the relevant regions' features. Supervised attention in the deep learning literature has been used to match a neural machine translations attention distribution to match an unsupervised aligner [16] and to match a sequence-to-sequence neural constituency parser's attention mechanism with traditional parsing features [17], for example. Our approach can be viewed as a form of supervised attention for document classification. The principle difference from existing work is that in supervised attention in the deep learning literature the method is trained in an end-to-end fashion with neural networks, whereas we train each module independently with classical methods and our feature representation for sentences are set of n-grams instead of dense real-valued vectors.

Rule-based line classifier

As a baseline, we also include a line classifier that selects relevant lines by searching for expert-generated keywords and phrases. After the lines are selected, the final representation is generated the same way, with the exception that all lines are given a weight of 1; thus, for all $l_i \varepsilon S_k$, $m(l_i) = 1$.

Oracle model

In addition to the line attention model, we also evaluate a model that uses the correct relevant lines from the annotator directly as input to the final classifier, which we refer to as the “oracle model”. Using the oracle lines, the final representation is generated the same way as the rule-based line classifier, where all lines are given a weight of 1.

3.7. Hyperparameter **tuning**

Similar to our baselines, we perform random search for 40 iterations and choose the hyperparameters that minimize 4-fold cross-validation error. The hyperparameters for our shallow attention method are an n-gram size for finding relevant lines between 1 and 4; an n-gram size for the second stage of making the final classification between 1 and 4.

For *xgboost*, the hyperparameters were 500 points from −2 to −0.5 in logspace for the learning rate; a max depth between 3 and 7; a minimum split loss reduction to split a node that is 0, 0.01, 0.05, 0.1, 0.5, or 1; a subsample ratio that is 0.5, 0.75, or 1; and an L2 regularization on the weights that is 0.1, 0.5, 1, 1.5, or 2.

For the final classifier, the L1 penalty is chosen from 500 evenly spaced points from −6 to 6 in logspace. Additionally, since the final representation is a weighted representation of the features of the top-k lines under the line classifier model, we have a hyperparameter k which determines how many lines to use, where k is between 1 and 5.

3.8. Ablation **experiments**

For our ablation experiments, we investigate the relative contribution of each component in our model.

No weighting

Here we investigate if weighting the features in each line by the classifier scores increases performance compared to weighting the features in each line by one.

No joining

Here we investigate how joining affects the results when information spans multiple lines. Instead of conjoining lines that occur adjacent to each other, we leave them as separate lines for our final classifier.

No weighting and no joining

Here we neither weight the features vectors representing each line nor do we join adjacent predicted lines.

3.9. Error analysis

To better understand model performance, we inspect all errors that the supervised line attention model makes for each attribute and cancer domain. In our investigation we find 6 primary types of errors, which we define below:

Attribute Qualification Error occurs when the model correctly extracts the relevant lines, but fails to classify the final label correctly because the label text is negated or qualified by an additional phrase indicating information is not available, such as in the following example: “If we were to classify the tumor, it would be grade 2 but due to the treatment effect it is unclassified.”

Rare Phrasing Error occurs when the model correctly predicts the relevant lines, but the relevant lines contain rare or unusual phrasing and the model assigns an incorrect final classification.

Irrelevant Lines Error occurs when the model includes irrelevant lines in its final predictions, which can influence the final classification.

Multi-Label Error occurs when a report contains a conjoined label (such as “grade 1 and grade 2”), but the model only correctly predicts one of the labels.

Annotator Error occurs when the model’s prediction is correct, but on re-review we noted that the annotator’s label was incorrect.

U*nknown error* occurs when the underlying cause of the error is not known. This often occurs when the model correctly extracts out the relevant line but assigns an incorrect final label.

4. Results

We trained our methods using various training set sizes of 32, 64, 128, and 186 with 4-fold cross validation. We take the average of 10 runs where we reshuffle the data and generate new splits each time and compute 95% confidence intervals for all methods using bootstrap resampling, with the exception of the HAN method due to computational limitations. For our results, unless stated otherwise, we are using the reduced annotation set due to its comparable annotation time to labeling the attribute values alone. As shown in Fig. 1 and Table 4, our shallow attention model frequently improves substantially over existing methods in terms of micro and macro-f1, particularly in the lowest data regimes. For example, for colon cancer we see an absolute improvement of 0.10 micro-f1 and 0.17 macro-f1 over previously existing methods with 32 labeled data points. Furthermore, SLA frequently tends to perform as well or better than state of the art methods with only half the labeled documents. Two exceptions are kidney cancer micro-f1 scores, where boosting performs 0.01 better in micro-f1. We see that the rule-based line classifier method tends to do better than existing methods with 64 labeled data points or fewer, but its performance plateaus and XGBoost outperforms it with 128 and 186 labeled data points. Furthermore, we see that the rule-based line classifier consistently performs worse than supervised line attention.

4.1. Ablation results

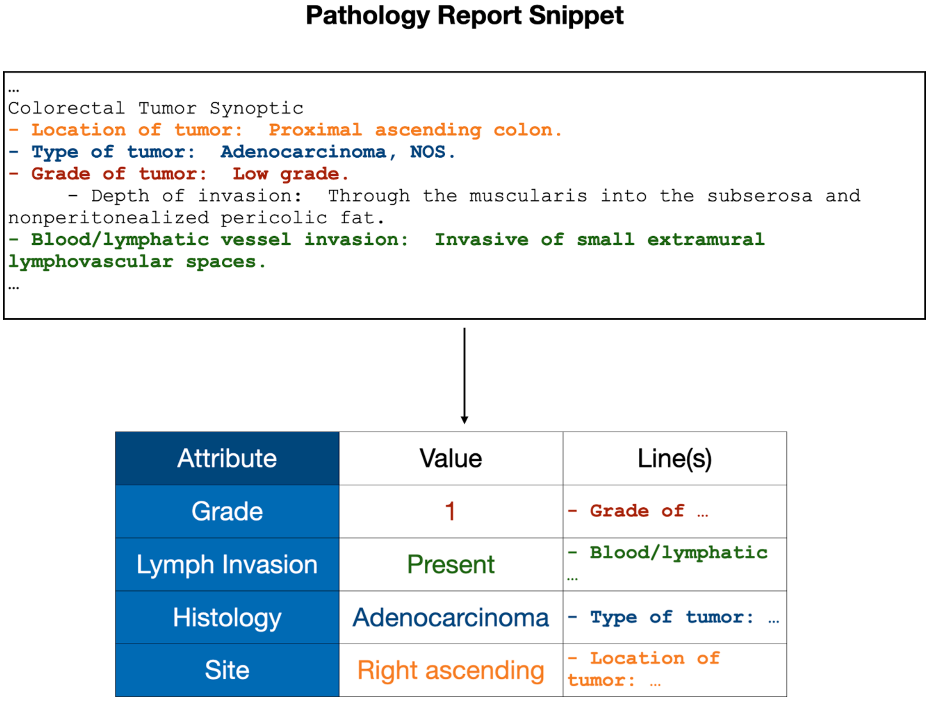
We plot the results of our ablation experiments in Fig. 2, using the same setup as our main result where we have training set sizes of 32, 64, 128, and 186 with 4-fold cross validation. Again, we take the average of 10 runs where we reshuffle the data and generate new splits each time and compute 95% confidence intervals for all methods using bootstrap resampling. We see mixed results for joining adjacent predicted lines; it appears to be inconsequential for colon cancer and detrimental for kidney cancer. However, weighting the features by line predictor seems beneficial for the macro-f1 scores. This seems to suggest that weighting helps primarily for rare classes since the macro-f1 score weights the f1 scores of each class equally.

4.2. Full annotations

Here we compare how well reduced annotation compares to the more laborious full annotation setting where we highlight all areas in the document relevant to final classification. We use the same setup as our main results and ablation experiments and present our results in Fig. 3. We can see that the full annotation set leads to a consistent increase in performance. However, it is unclear whether the extra time required to create this full annotation scheme is beneficial overall as it would lead to fewer documents annotated in the same amount of time.

4.3. Error analysis

We provide a compilation of the number of errors across attributes in Table 2 and Table 3 for colon and kidney cancer, respectively. We see that the most common error is the multi-label error. This is primarily problematic for colon cancer histologic grade, where pathologists will describe a range of grades such as “grade 1–2” and tumor site for colon and kidney cancer as tumors can inhabit multiple sites. This suggests that treating this as a multi-label classification problem instead of

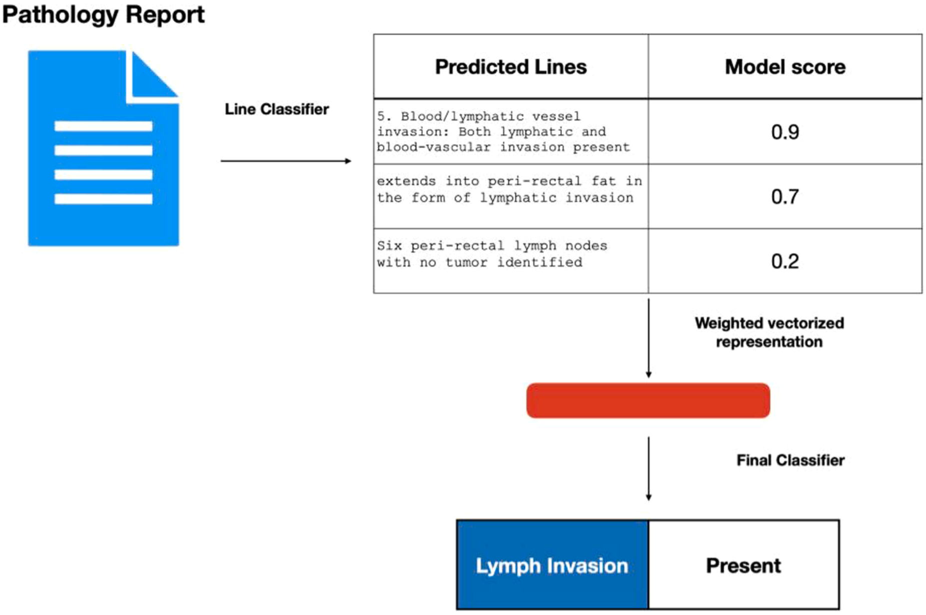


Pathology Report Snippet

```
…
Colorectal Tumor Synoptic
- Location of tumor:  Proximal ascending colon.
- Type of tumor:  Adenocarcinoma, NOS.
- Grade of tumor:  Low grade.
    - Depth of invasion:  Through the muscularis into the subserosa and
nonperitonealized pericolic fat.
- Blood/lymphatic vessel invasion:  Invasive of small extramural
lymphovascular spaces.
…
```

Attribute	Value	Line(s)
Grade	1	- Grade of …
Lymph Invasion	Present	- Blood/lymphatic …
Histology	Adenocarcinoma	- Type of tumor: …
Site	Right ascending	- Location of tumor: …

Fig. 1. Average micro-f1 and macro-f1 performance across attributes of different methods as a function of 32, 64, 128, 186 labeled examples on colon cancer and kidney cancer pathology reports. SLA: supervised line attention; oracle: oracle model that gets access to the true lines as input; rules: line prediction done with a rule-based method and logistic regression to predict the final class; boost: XGBoost; SVM: Support Vector Machine; logistic; logistic regression; RF: Random forest; HAN: Hierarchical attention network. We present the mean result across 10 random shufflings of the data as well as 95% bootstrap confidence intervals. We see that our method SLA outperforms existing methods in almost all cases. Furthermore, we see that predicting relevant lines outperforms our rule-based method to extract relevant lines.



Predicted Lines	Model score
5. Blood/lymphatic vessel invasion: Both lymphatic and blood-vascular invasion present	0.9
extends into peri-rectal fat in the form of lymphatic invasion	0.7
Six peri-rectal lymph nodes with no tumor identified	0.2

Lymph Invasion	Present

Fig. 2. Ablation studies for SLA measuring the average micro-f1 and macro-f1 performance across attributes of different methods as a function of 32, 64, 128, 186 labeled examples on colon cancer and kidney cancer pathology reports. We investigate the impact of joining adjacent selected lines prior to featurization as well as the impact of weighting the features by the line classifier scores. We present the mean result across 10 random shufflings of the data with 95% bootstrap confidence intervals. While it appears that joining adjacent predicted lines leads to mixed or potentially even negative performance over not joining adjacent predicted lines, weighted methods seem to outperform their unweighted alternatives, especially for macro-f1 scores, suggesting that weighting helps in particular for rare classes.

naively conjoining multiple labels may reduce many of the errors.

5. Discussion

We have investigated the efficacy of location-enriched annotations and a corresponding hierarchical method, which we call Supervised Line Attention, for extracting data elements from pathology reports across colon and kidney cancers at UCSF. By leveraging location annotations, our two-stage modeling approach can lead to increases of micro-f1 scores up to 0.1 and macro-f1 scores up to 0.17 over state-of-the-art methods and typically reduces the number of training data points by 50 percent to achieve performance of existing methods.

Our hierarchical modeling approach with enriched annotations was primarily developed to tackle the problem of efficiently achieving accurate performance. Previous approaches that attempt to leverage additional data use multi-task learning and transfer learning using information from other cancer domains with complex modeling architectures. For example, Qui et al. investigated using transfer learning with convolutional neural networks to extract data from 942 breast and lung cancer reports, achieving 0.685 and 0.782 micro-f1 scores, respectively. [24] Alawad *et al.* implemented multitask learning with convolutional neural networks to classify tumor attributes in 942 pathology reports for breast and lung cancers, and achieved 0.77, 0.79, and 0.96 micro-f1 scores for tumor site, histologic grade, and laterality, respectively. [25]

An important observation is that our hierarchical approach is more

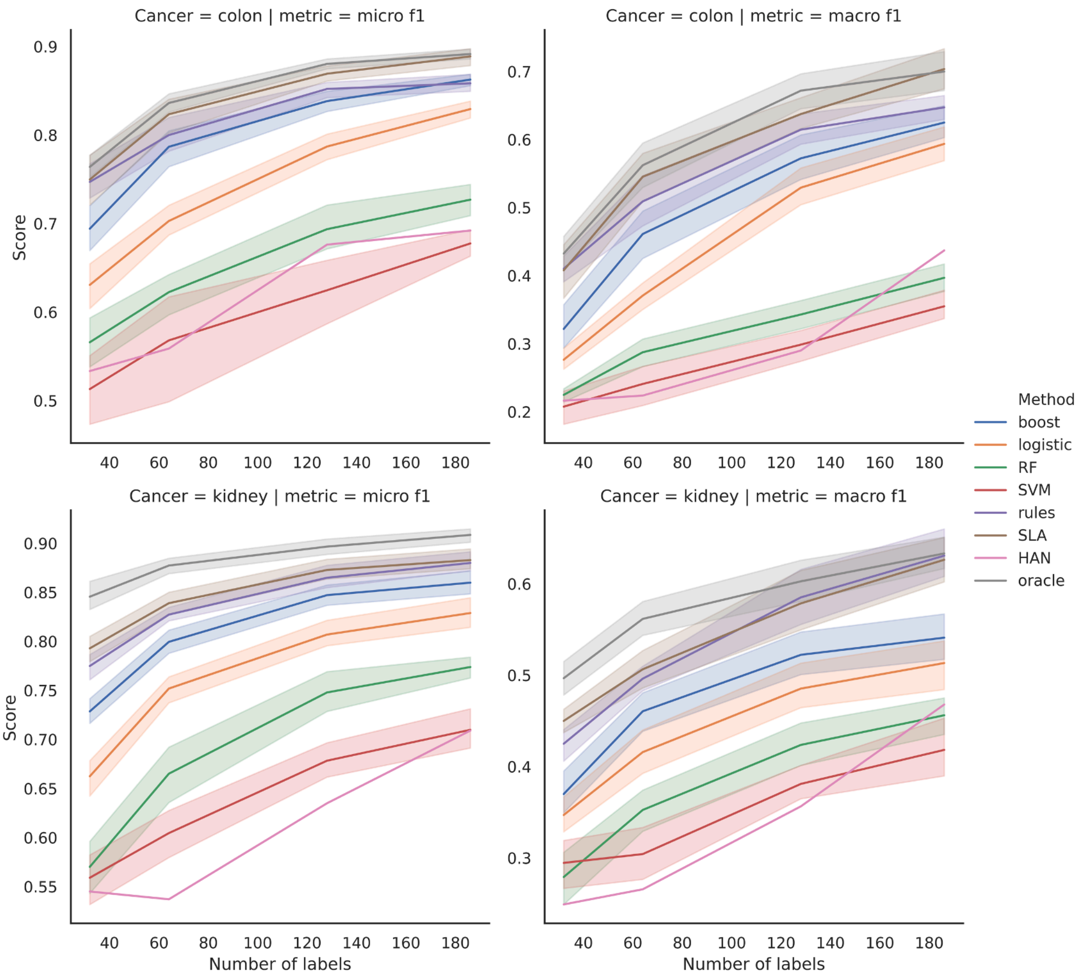


Fig. 3. Comparing the more laborious annotating the location information all relevant lines for a given attribute as compared to the more lightweight annotation method of only annotating the first line in the synoptic comment, if the synoptic comment contains the information, or the first relevant line in the document otherwise. We see that having the additional information yields a consistent, though sometimes small, benefit.

Table 2
Error analysis: Colon cancer.

Attribute	Histologic Grade	Histologic Type	Perineural invasion	Lymphovascular invasion	Procedure	Tumor Site	Total
Attribute Qualification Error	1	0	0	0	0	0	1
Rare phrasing	0	0	0	**1**	3	0	4
Irrelevant Lines Error	1	0	0	0	5	0	6
Annotator Error Error	3	**1**	**1**	0	5	0	10
Multi-label Error	**6**	0	0	0	0	**6**	**12**
Unknown error	1	0	0	0	**6**	0	7
Total by attribute	12	1	1	1	**19**	6	40

Table 3
Error analysis: Kidney cancer.

Attribute	Histologic Grade	Histologic Type	Specimen Laterality	Lymphovascular invasion	Procedure	Tumor Site	Total
Attribute Qualification Error	0	0	0	0	0	0	0
Rarephrasing	0	0	0	**1**	**5**	0	6
Irrelevant Lines Error	**1**	0	0	**1**	1	1	4
Annotator Error	**1**	2	0	**1**	1	0	5
Multi-label Error	0	**4**	0	0	2	**6**	**12**
Unknown error	**1**	**4**	0	**1**	1	5	**12**
Total by attribute	3	10	0	4	10	**12**	39

interpretable than previous machine learning methods, since addition to outputting the probability and predicted value for a certain report, our system outputs the exact lines of the text used to make the classification as well. This enables practitioners to easily check predictions by examining the lines output by the extraction system, and verify the system is working as expected before making clinical decisions. The hierarchical

Table 4
Average micro-f1 and macro-f1 performance across attributes of different methods as a function of 32, 64, 128, 186 labeled examples on colon and kidney cancer. Highest performing non-oracle method is bolded.

	HAN	RF	SVM	Boost	Logistic	Rules	SLA	Oracle
Colon								
Micro-F1								
32	0.45	0.57	0.51	0.69	0.63	0.75	**0.80**	0.80
64	0.50	0.62	0.57	0.79	0.70	0.79	**0.84**	0.84
128	0.60	0.69	0.62	0.84	0.79	0.83	**0.87**	0.88
186	0.61	0.73	0.68	0.86	0.83	0.84	**0.89**	0.89
Macro-F1								
32	0.18	0.22	0.21	0.32	0.28	0.43	**0.50**	0.50
64	0.19	0.29	0.24	0.46	0.37	0.52	**0.60**	0.59
128	0.25	0.34	0.30	0.57	0.53	0.56	**0.66**	0.67
186	0.37	0.40	0.35	0.62	0.59	0.59	**0.69**	0.70
Kidney								
Micro-F1								
32	0.54	0.57	0.56	0.73	0.66	0.79	**0.80**	0.85
64	0.54	0.67	0.60	0.80	0.75	0.81	**0.83**	0.86
128	0.63	0.75	0.68	**0.85**	0.81	0.83	0.84	0.87
186	0.71	0.77	0.71	**0.86**	0.83	0.84	0.85	0.88
Macro-F1								
32	0.25	0.28	0.29	0.37	0.35	0.46	**0.48**	0.51
64	0.27	0.35	0.30	0.46	0.42	0.49	**0.52**	0.54
128	0.36	0.42	0.38	0.52	0.49	0.51	**0.54**	0.55
186	0.47	0.46	0.42	0.54	0.51	0.54	**0.56**	0.56

attention approach used by Gao *et al.* also can output the most pertinent sentences for a classification by using the attention mechanism to hierarchically filter out pieces of text. [10] However, our experiments show that HAN requires a large training size to achieve adequate performance due to the more complex architecture used, and requires significantly more development and computational time to search the hyperparameter space. Additionally, there have been recent concerns regarding the interpretability of attention distributions from neural networks. [18]

Our study has a few limitations. Although we demonstrate the high performance of our methodology is applicable to both colon and kidney cancer, our investigation was done at a single institution; this may limit the generalizability of our findings to other institutions that use different pathology reporting or data collection systems. Second, within the field of natural language processing, there has been strong empirical evidence showing the benefit of pre-trained contextualized representations for a variety of tasks, both in and out of clinical applications [19–22]. In preliminary experiments, we investigated the efficacy of using biomedical word vectors [23] as feature representation input to our SLA model, but did not see an improvement in results. However, it would be interesting to investigate the effect that more sophisticated contextualized representations may have on downstream performance.

6. Conclusion

We have shown that including location information in annotation and applying our supervised line attention mechanism can vastly reduce the number of labeled documents needed for accurate tumor attribute classification compared to state of the art approaches. Furthermore, our supervised line attention method allows for greater interpretability due to its hierarchical nature, which can allow for easy verification of its outputs for clinicians. We hope these methods will advance the application of information extraction in medicine, where labeled data is scarce and expensive to acquire.

Funding

Bakar Computational Health Sciences Institute, University of California, San Francisco.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Partial support is gratefully acknowledged from ARO grant W911NF1710005, NSF grants DMS-1613002 and IIS 1741340, the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370, Adobe Research Grant, and the Bakar Computational Health Sciences Institute, University of California, San Francisco.

References

[1] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington (DC): : National Academies Press (US) 2011. http://www.ncbi.nlm.nih.gov/books/NBK91503/ (accessed 6 May 2020).

[2] G. Burger, A. Abu-Hanna, N. de Keizer, et al., Natural language processing in pathology: a scoping review. J. Clin. Pathol. Published Online First: 22 July 2016. http://doi.org/10.1136/jclinpath-2016-203872.

[3] A. Yala, R. Barzilay, L. Salama, M. Griffin, G. Sollender, A. Bardia, C. Lehman, J. M. Buckley, S.B. Coopey, F. Polubriaginof, J.E. Garber, B.L. Smith, M.A. Gadd, M. C. Specht, T.M. Gudewicz, A.J. Guidi, A. Taghian, K.S. Hughes, Using machine learning to parse breast pathology reports, Breast Cancer Res. Treat. 161 (2) (2017) 203–211, https://doi.org/10.1007/s10549-016-4035-1.

[4] V. Jouhet, G. Defossez, A. Burgun, P. le Beux, P. Levillain, P. Ingrand, V. Claveau, Automated classification of free-text pathology reports for registration of incident cases of cancer, MethodsInf. Med. 51 (03) (2012) 242–251, https://doi.org/10.3414/ME11-01-0005.

[5] J.X. Qiu, H.-J. Yoon, P.A. Fearn, G.D. Tourassi, Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports, IEEE J. Biomed. Health Inform. 22 (1) (2018) 244–251, https://doi.org/10.1109/JBHI.622102010.1109/JBHI.2017.2700722.

[6] A. Odisho*, B. Park, N. Altieri, et al. PD58-09 EXTRACTING STRUCTURED INFORMATION FROM PATHOLOGY REPORTS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING. J Urol Published Online First: April 2019.https://www.auajournals.org/doi/abs/10.1097/01.JU.0000557177.97226.63 (accessed 23 Feb 2020).

[7] Aspect-augmented Adversarial Networks for Domain Adaptation | Transactions of the Association for Computational Linguistics | MIT Press Journals. https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_0007(accessed 23 Feb 2020).

[8] Stubbs A. MAE and MAI: Lightweight Annotation and Adjudication Tools. 2011. 129–33.https://www.aclweb.org/anthology/W11-0416 (accessed 24 Feb 2020).

[9] Cancer Protocol Templates. Coll. Am. Pathol. https://www.cap.org/protocols-and-guidelines/cancer-reporting-tools/cancer-protocol-templates (accessed 6 May 2020).

[10] S. Gao, M.T. Young, J.X. Qiu, et al. Hierarchical attention networks for information extraction from cancer pathology reports, J. Am. Med. Inform. Assoc. 25 (2018) 321–330. http://doi.org/doi:10.1093/jamia/ocx131.

[11] J. Bergstra, Y. Bengio, Random Search for Hyper-Parameter Optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[12] PedregosaFabian, VaroquauxGaël, GramfortAlexandre, et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. Published Online First: 1 November 2011. https://dl.acm.org/doi/abs/10.5555/1953048.2078195 (accessed 24 Feb 2020).

[13] K. Cho, B. van Merriënboer, C. Gulcehre, et al., Learning Phrase Representations using RNN Encoder–Decoder for Statistical, Mach. Transl. (2014) 1724–1734, https://doi.org/10.3115/v1/D14-1179.

[14] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. Published Online First: 22 December 2014. https://arxiv.org/abs/1412.6980v9 (accessed 24 Feb 2020).

[15] XGBoost: A Scalable Tree Boosting System. https://www.kdd.org/kdd2016/subtopic/view/xgboost-a-scalable-tree-boosting-system/670 (accessed 24 Feb 2020).

[16] L. Liu, M. Utiyama, A. Finch, et al. Neural Machine Translation with Supervised Attention. 2016. 3093–102. https://www.aclweb.org/anthology/C16-1291 (accessed 23 Feb 2020).

[17] H. Kamigaito, K. Hayashi, T. Hirao, et al. Supervised Attention for Sequence-to-Sequence Constituency Parsing. 2017. 7–12. https://www.aclweb.org/anthology/I17-2002 (accessed 23 Feb 2020).

[18] S. Jain, B.C. Wallace, Attention is not Explanation. ArXiv190210186 Cs Published Online First: 8 May 2019. http://arxiv.org/abs/1902.10186 (accessed 23 Apr 2020).

[19] M.E. Peters, M. Neumann, M. Iyyer, et al. Deep contextualized word representations. Published Online First: 15 February 2018. https://arxiv.org/abs/1802.05365v2 (accessed 20 Jun 2020).

[20] J. Devlin, M.-W. Chang, K. Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published Online First: 11 October 2018. https://arxiv.org/abs/1810.04805v2 (accessed Jun 2020).

[21] J. Lee, W. Yoon, S. Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Published Online First: 25 January 2019. http://doi.org/10.1093/bioinformatics/btz682.

[22] K. Huang, J. Altosaar, R. Ranganath, ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Published Online First: 10 April 2019. https://arxiv.org/abs/1904.05342v2 (accessed 20 Jun 2020).

[23] S. Pyysalo, F. Ginter, H. Moen, et al., Distributional Semantics Resources for Biomedical Text Processing, 5.

[24] J. Qui, H. Yoon, P. Fearn, G. Tourassi, Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports, IEEE J. Biomed. Health. Inf. (2017).

[25] M. Alaward, H. Yoon, G. Tourassi, Coarse-to-Fine Multi-Task Training of Convolutional Neural Networks for Automated Information Extraction from Cancer Pathology Reports. 2018. 4-7. IEEE EMBS International Conference on Biomedical and Health Informatics.