Title

Large Scale Semi-Automated Labeling of Routine Free-Text Clinical Records for Deep Learning

Permalink

https://escholarship.org/uc/item/7005m332

Journal

Journal of Digital Imaging, 32(1)

ISSN

0897-1889

Authors

Trivedi, Hari M
Panahiazar, Maryam
Liang, April
et al.

Publication Date

2019-02-01

DOI

10.1007/s10278-018-0105-8

Peer reviewed

Journal of Digital Imaging (2019) 32:30–37
https://doi.org/10.1007/s10278-018-0105-8

Large Scale Semi-Automated Labeling of Routine Free-Text Clinical Records for Deep Learning

Hari M. Trivedi[1] · Maryam Panahiazar[2] · April Liang[3] · Dmytro Lituiev[2] · Peter Chang[1] · Jae Ho Sohn[1] · Yunn-Yi Chen[4] · Benjamin L. Franc[1] · Bonnie Joe[1] · Dexter Hadley[2]

Published online: 20 August 2018

Abstract

Breast cancer is a leading cause of cancer death among women in the USA. Screening mammography is effective in reducing mortality, but has a high rate of unnecessary recalls and biopsies. While deep learning can be applied to mammography, large-scale labeled datasets, which are difficult to obtain, are required. We aim to remove many barriers of dataset development by automatically harvesting data from existing clinical records using a hybrid framework combining traditional NLP and IBM Watson. An expert reviewer manually annotated 3521 breast pathology reports with one of four outcomes: left positive, right positive, bilateral positive, negative. Traditional NLP techniques using seven different machine learning classifiers were compared to IBM Watson's automated natural language classifier. Techniques were evaluated using precision, recall, and F-measure. Logistic regression outperformed all other traditional machine learning classifiers and was used for subsequent comparisons. Both traditional NLP and Watson's NLC performed well for cases under 1024 characters with weighted average F-measures above 0.96 across all classes. Performance of traditional NLP was lower for cases over 1024 characters with an F-measure of 0.83. We demonstrate a hybrid framework using traditional NLP techniques combined with IBM Watson to annotate over 10,000 breast pathology reports for development of a large-scale database to be used for deep learning in mammography. Our work shows that traditional NLP and IBM Watson perform extremely well for cases under 1024 characters and can accelerate the rate of data annotation.

Keywords IBM Watson · Machine learning · Artificial intelligence · Deep learning · Natural language processing (NLP) · Pathology · Mammography

Introduction

Breast cancer is one of the leading causes of cancer death among women in the USA [1, 2], and more than 310,000 new cases will be diagnosed in 2017 [3]. Screening mammography has proven to be an effective tool for reducing breast cancer mortality by allowing early detection of suspicious findings such as masses, abnormal calcifications, architectural distortion, and asymmetries [4]. Sensitivity is reported upwards of 85% [5]; however, this is accompanied by a high proportion of "recall" imaging for further evaluation of potentially suspicious findings. To illustrate, per 1000 women who receive annual mammographic screening, approximately 80 are required to return for recall imaging, 30 must undergo biopsy, and this ultimately results in the detection of only eight cancers [6]. Therefore, there is significant room for improvement in reducing unnecessary recall imaging and biopsies.

Deep learning has recently demonstrated exceptional performance in medical image recognition tasks [7], including detection of breast and prostate cancers on histopathology slides [8], diagnosis of Alzheimer's disease from MRI/PET imaging [9], and classification of skin cancer based on lesion photographs [10]. Deep learning has also been applied to mammography [11–15], but development of high-performance algorithms

✉ Hari M. Trivedi
hari.trivedi@gmail.com

[1] Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA

[2] Institute for Computational Health Sciences, University of California, San Francisco, CA, USA

[3] University of California School of Medicine, San Francisco, CA, USA

[4] Department of Pathology, University of California, San Francisco, CA, USA

requires extremely large, well-annotated datasets for training. A typical dataset consists of medical images annotated with ground-truth labels such as pathologic outcome or clinical diagnosis. The deep learning algorithm then trains itself to recognize underlying differences between negative and positive cases. Unfortunately, the majority of medical imaging datasets have only several hundreds or thousands of training examples, and oftentimes are heavily imbalanced towards negative or benign cases [16]. For comparison, the first deep learning model to surpass human-level performance on general image classification tasks trained on nearly 1.2 million images [17].

Development of large, well-annotated datasets is hindered by several factors including lack of funding, prohibitive requirements in time and medical expertise, and privacy issues that complicate sharing [16]. For this reason, development of these datasets has traditionally required manual efforts from a large team (such as a through a clinical trial). However, the sheer number of cases required for effective deep learning makes these types of manual methods unfeasible, if not impossible.

We postulate that the burden of dataset construction can be significantly reduced by automating the structuring and annotation of existing routine clinical records. However, the majority of relevant clinical information is stored as free-text, making extraction into a structured format laborious and expensive. At least part of this difficulty arises from a high level of noise in the data such as misspellings, abbreviations, acronyms, poor grammatical structure, and variations in reporting styles which makes automatic interpretation challenging [18]. Many of these issues are addressed by existing natural language processing and machine learning frameworks such as cTakes and WEKA [19–24], but their implementation requires significant domain expertise and programming knowledge. In recent years, however, newer automated solutions such as IBM Watson have been developed to provide the ability to perform text classification with little domain knowledge [25–27].

In this paper, we present a semi-automated framework that combines and compares traditional natural language processing techniques (traditional NLP) with the proprietary IBM Watson Natural Language Classifier (Watson NLC). Our aim was to label more than 10,000 free-text breast pathology reports with the final pathologic diagnosis to serve as ground truth for annotating mammographic images.

Materials and Methods

Patient Characteristics/Study Cohort

Pertaining to 7237 women (mean age = 51.8 years), 10,420 reports from 1997 to 2014 were extracted from an in-house pathology database. Report types included all breast specimens, including fine needle aspirations (FNA), core biopsies, lumpectomies, and mastectomies. Only the “final diagnosis” section (inclusive of any addenda) was considered in analysis; additional fields such as clinical history and comments were not utilized. Due to input length limitations in IBM Watson’s natural language classifier, reports over 1024 characters ($n = 522$) were separated for later analysis (long set), resulting in 9898 remaining reports under 1024 characters. From these, a random sample of 3099 reports was selected (standard set). Both the standard set and long set were manually labeled by an expert reviewer with the aid of a board-certified pathologist. Ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), and metastatic nodal disease were considered positive. All other findings, including lobular carcinoma in situ (LCIS) or pre-malignant lesions such as radial scar, were considered negative. Indeterminate lesions or insufficient samples were also considered negative as these were typically followed up by more conclusive sampling.

In order to devise a labeling structure appropriate for all reports, the reports were first divided by the laterality of the reported specimen: left, right, or bilateral. Unilateral reports could be positive (left positive, right positive) or negative. Bilateral reports could be positive for a single breast (left positive, right positive), positive for both breasts (bilateral positive), or negative for both breasts (negative). Thus, a four-class labeling system could be utilized for all reports: negative, left positive, right positive, and bilateral positive. The distribution of cases for the standard and long sets is shown in Table 1.

Table 1 Distribution of cases in the standard set (< 1024 characters) and long set (> 1024 characters) by class

Report laterality	Class	Short set	Long set
Bilateral	Bilateral positive	13	81
	Left positive	43	112
	Right positive	40	35
	Negative	232	128
	Total	328	356
Left	Left positive	450	92
	Negative	954	4
	Total	1404	96
Right	Right positive	421	70
	Negative	855	0
	Total	1276	70
Not Specified	Negative	91	0
	Total	*91*	*0*
Grand Total		3099	522

Using this labeling scheme, we developed an annotation framework using a combination of traditional NLP and IBM Watson's NLC, an overview of which is shown in Fig. 1.

Traditional Natural Language Processing

For traditional natural language processing and classification, a standard pipeline was used consisting of three steps: preprocessing, text mining, and classification.

Text Mining: Pre-Processing

The main challenge with text mining in the clinical domain is the high level of noise in the corpus and training data. This is largely due to unknown words (misspellings, medical terminology, or acronyms), non-words (punctuation, numbers), common prepositions, and sentence fragments. For the purposes of this study, all non-words were removed. Sentence fragments and prepositions are addressed by tokenization and term frequency-inverse document frequency (TF-IDF), as described below.

Text Mining: Tokenization

The pathology report text was converted to the ARFF file format and imported to WEKA—a machine learning framework which can be used for NLP [23]. Each report was first tokenized, which is defined as demarcation of discrete sections within a string. In our case, each token was an N-gram (a set of co-occurring words) in lengths of one to three words. For example, the phrase "no ductal carcinoma" will result in the following tokens: "no," "ductal," "carcinoma," "no ductal," "ductal carcinoma," and "no ductal carcinoma." Each pathology report was thusly tokenized into a vector of N-grams to serve as the input for TF-IDF.

Text Mining: Term Frequency-Inverse Document Frequency

TF-IDF was used to assign importance to individual tokens within a report. Term frequency (TF) is defined as the number of times a token appears in the report, which serves as an estimate of its importance. However, this leads to common but irrelevant tokens such as "of" and "the" being assigned high importance. To combat this, inverse document frequency (IDF) is calculated based on the uniqueness of the token in the overall corpus. The product of the TF and IDF assigns a final, weighted importance to each token.

Classification Using Machine Learning

Following construction of the TF-IDF matrix, seven supervised machine learning algorithms were tested to determine the best performing classifier for predicting the label for each report: PART, decision tables, AdaBoost, Naive Bayes, multiclass logistic regression (one vs. all), support vector machine (SVM), and majority vote classifier (ZeroR). For our dataset, logistic regression outperformed all other classifiers (Fig. 2) and was chosen as the classifier for all subsequent steps.

Watson Natural Language Classifier

The IBM Watson Natural Language Classifier was accessed via the IBM Bluemix online portal for training and the application program interface (API) for testing. Training data was uploaded to the online portal as a spreadsheet containing the free-text reports in one column and the corresponding ground truth label in the second column. A classifier was then automatically generated by training on this data. Test cases were uploaded in batch through the API which returned the top

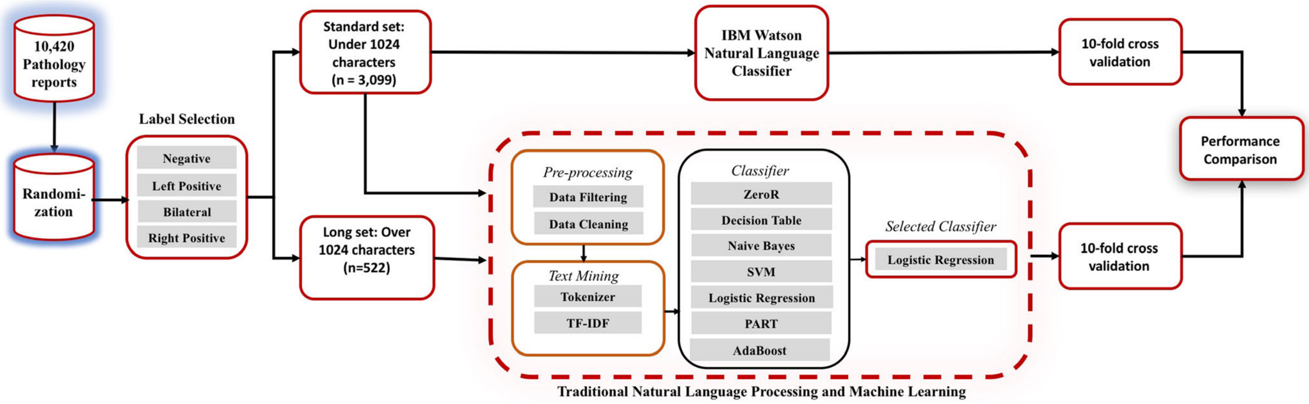

Fig. 1 Overview of the annotation pipeline demonstrating both the traditional NLP and IBM Watson arms. Reports were divided by length, as reports over 1024 characters (long set) could not be processed by Watson. Reports under 1024 characters (standard set) were processed using both traditional NLP and Watson's NLC. Logistic regression was the top-performing classified in preliminary testing of traditional NLP and was therefore used for all subsequent comparisons

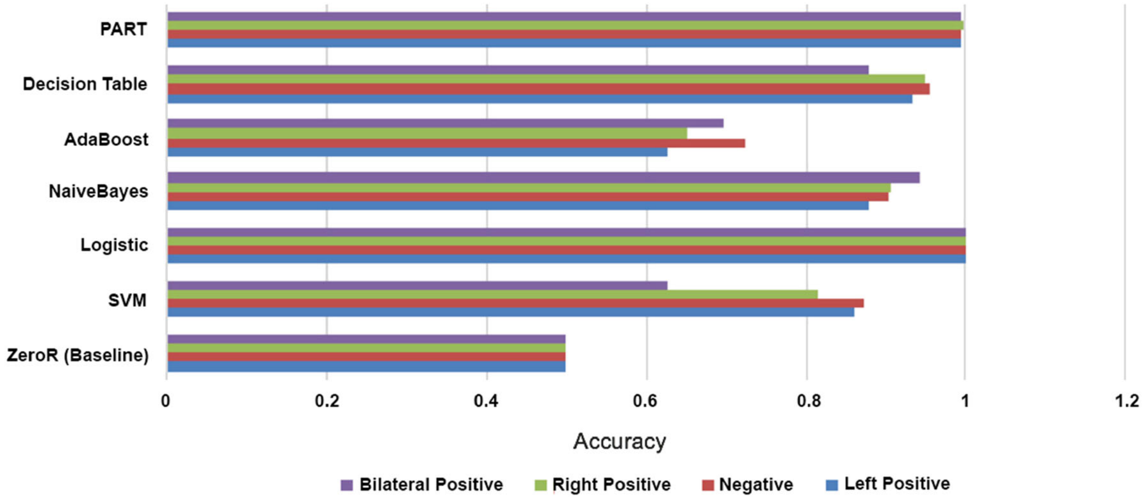

Fig. 2 Accuracy of various machine learning classifiers for traditional NLP. Logistic regression achieved the best performance and was thus selected for all subsequent comparisons

three class predictions and the corresponding confidence scores for each.

Experimental Setup and Evaluation Framework

Standard Set

The 3099 cases were randomly divided into 80% training/20% test sets 10 times for performance of 10-fold cross validation, resulting in 2480 training and 619 test cases in each fold. Both traditional NLP and Watson NLC were trained and tested on each fold. The performance of each classifier was averaged across all 10 folds to obtain a representative example of its performance. Precision, recall, and F-measure were calculated for each class, as well as the weighted averages of these metrics across all four classes.

Long Set

All 522 cases in the long set were evaluated using traditional NLP. Watson NLC could not be used due to its character limitation. These cases were similarly randomly divided into 80% training/20% test sets 10 times for performance of 10-fold cross validation, resulting in 420 training and 102 test cases in each fold. The output from the traditional NLP classifier was recorded for each test case. Precision, recall, and F-measure were calculated for each class, as well as the weighted averages across all four classes.

Results

The mean length of each report across the entire dataset was 408.4 characters. The most common words in the corpus are shown in Fig. 3.

Standard Set

The mean length of cases in this category was 239.2 characters. Across all 10 folds, the mean number of the left positive, right positive, bilateral positive, and negative cases in the test set were 100.3, 88.9, 2.5, and 427.3, respectively. Both traditional NLP and Watson NLC exhibited excellent performance for the left positive, right positive, and negative cases, with average F-measures greater than 0.9 (Fig. 4; Table 2). Watson NLC demonstrated a slight performance advantage over traditional NLP. The best performance was achieved in negative cases, with F-measures around 0.99 for both traditional NLP and Watson NLC. The worst performing category was the bilateral positive cases with F-measures of 0.1 and 0.05 for traditional NLP and Watson NLC, respectively. When considering weighted averages across all classes, both traditional NLP and Watson NLC performed extremely well with F-measures above 0.96 (Fig. 5).

Long Set

As previously mentioned, due to Watson NLC's inherent character limit, only traditional NLP was used to classify cases over 1024 characters. The average length of cases in this category was 1412.9 characters with a maximum length of 4586 characters. Across all 10 folds, the average number of left positive, right positive, bilateral positive, and negative cases in the test set was 39.6, 39.5, 15.8, and 7.1, respectively. Classifier performance was lower for left positive, right positive, and negative cases as compared to cases under 1024 characters (Fig. 4; Table 2). However, the performance for bilateral positive cases was considerably higher. Furthermore, the vast majority of classification errors in the bilateral positive class consisted of assigning left or right positive labels, rather than a negative label. The weighted average across all classes for F-measure across was 0.83, lower than that of cases in the standard set (Fig. 5).

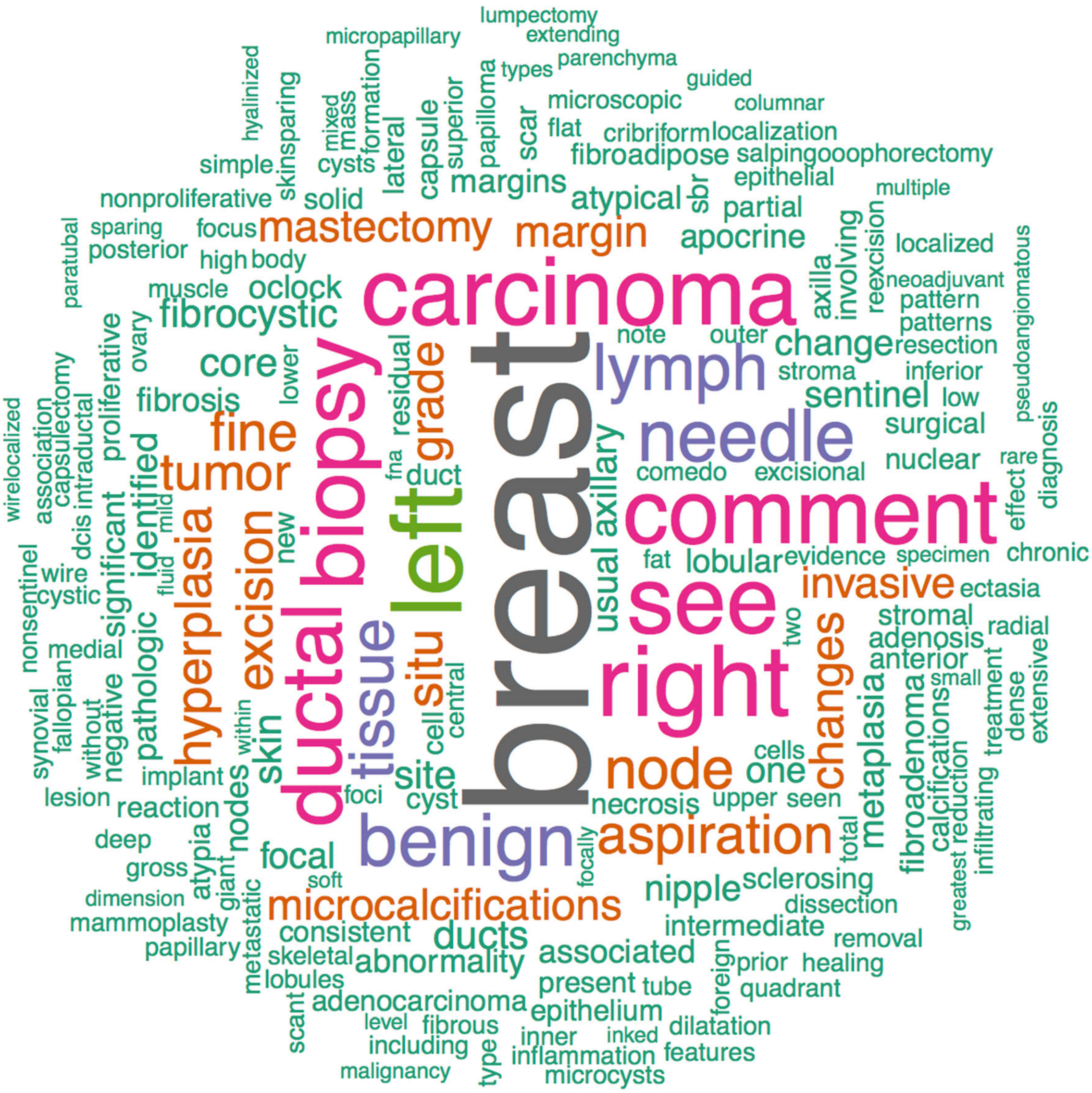

Fig. 3 Word cloud demonstrating the most common words in our dataset of breast pathology reports

Fig. 4 F-measures demonstrating performance of traditional NLP and Watson's NLC for each class. Performance for bilateral positive cases was poor for the standard set, likely due to the small number of training and test cases. Performance improved considerably for bilateral positive cases in the long set, but was slightly worse for the remaining classes, likely due to increased complexity of the longer reports. For the standard set, performance between Watson's NLC and traditional NLP was comparable

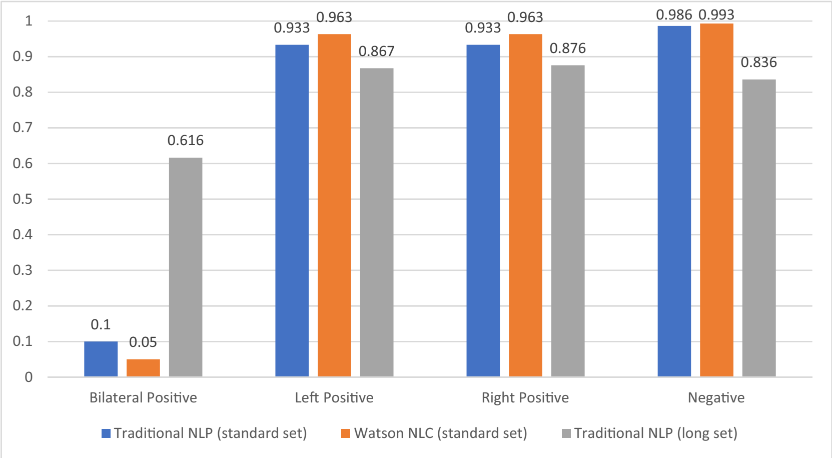

Table 2 Detailed performance statistics by class for traditional NLP and Watson NLC classifiers on the standard set and long sets

		Traditional NLP (standard set)	Watson NLC (standard set)	Traditional NLP (long set)
Precision	Bilateral	0.100	0.050	0.726
	Negative	0.980	0.991	0.812
	Left positive	0.942	0.963	0.842
	Right positive	0.941	0.953	0.862
	Weighted average	0.965	0.977	0.837
Recall	Bilateral	0.100	0.050	0.548
	Negative	0.992	0.995	0.876
	Left positive	0.925	0.963	0.898
	Right positive	0.925	0.953	0.892
	Weighted average	0.968	0.980	0.837
F-measure	Bilateral	0.100	0.050	0.616
	Negative	0.986	0.993	0.836
	Left positive	0.933	0.963	0.867
	Right positive	0.933	0.953	0.876
	Weighted average	0.966	0.979	0.832

Discussion

In this work, we demonstrate the feasibility of creating a hybrid framework using traditional NLP and automated solutions from IBM Watson to derive pathologic diagnosis from free-text breast pathology reports. This technique significantly accelerates the rate of extraction of meaningful data from clinical free-text reports and has important implications for improving the quantity and quality of large-scale datasets available for deep learning. To our knowledge, no such application for processing of free-text pathology records has yet been described.

Both traditional NLP and Watson's NLC performed well across all classes, particularly for cases in the standard set for which both classifiers achieved overall F-measures over 0.96. It is also worth noting that both techniques performed well despite the innate high variability in the dataset which included reports over a 17-year span, containing many different specimen types, cancer types, verbiage, and the reporting styles of dozens of pathologists.

The only class in which both classifiers performed poorly was the bilateral positive class. This can be attributed to the paucity of training examples, particularly in the standard set. There was a mean of only 11.5 training examples per fold in the standard set, representing less than 0.5% of cases. For comparison, there was a mean of 65.2 training examples per fold in the long set, and classification performance in this category was markedly improved. Nevertheless, the small sample size of test cases in either set makes these results difficult to interpret. It is also worth reiterating that the character limit of IBM Watson prevented its application to cases over 1024 characters, which represented approximately 8% of our

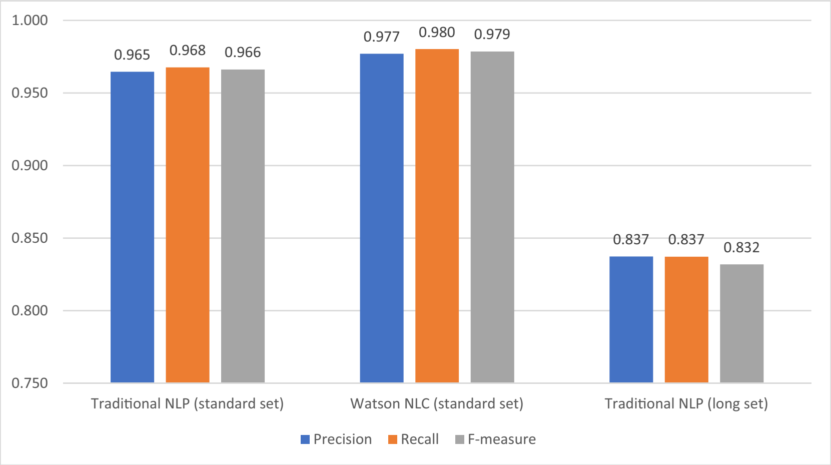

Fig. 5 Weighted averages for precision, recall, and F-measure of each classifier across all four classes. Overall performance of Watson NLC and traditional NLP was excellent for the standard set. Overall performance of traditional NLP for the long set was slightly worse, likely owing to increased complexity of the longer reports

dataset. Truncation of longer reports was considered; however, the longest reports were disproportionately for bilateral breasts which were reported in an arbitrary structure. For this reason, we believe truncation would generate unpredictable, and thus ultimately unreliable, results. To address both of these issues and improve overall accuracy, future work includes chunking reports into individually labeled specimens that should fall within Watson character limitations while also providing a less noisy and thus more robust training set.

Despite these limitations, we demonstrate Watson's automated natural language classifier provides a powerful tool for interpreting medical pathology reports for breast cancer outcomes. While traditional natural language processing techniques can be powerful, their application requires technical knowledge and programming experience to create a semi-automated pipeline and iterative models must be evaluated to produce optimal results. Conversely, "black-box" proprietary solutions, such as IBM Watson, require simply submitting a labeled spreadsheet to a web portal to achieve reassuringly comparable performance to traditional NLP techniques. For clinicians and other annotators that may be devoid of an NLP programming knowledge base, Watson and other black-box methods can serve as a pragmatic "litmus test" of a dataset to determine whether signal exists before devoting further resources. However, depending on the used case, these benefits may be outweighed by the cost and black-box nature of a proprietary system with no ability to modify the algorithm to improve performance.

The overall utility of these results for construction of a deep learning database of labeled mammograms is worth considering. Independently, errors of 3–4% for the standard set and 17% for the long set may be considered unacceptable. However, in our experience, nearly all positive cases of breast cancer resulted in more than one positive pathology report (i.e., FNA followed by core biopsy or lumpectomy). Because only one positive pathology result is required to label a mammogram as positive, almost all false-negative results were superseded by a subsequent true positive result, thus resulting in the mammogram being accurately labeled as positive. To validate this, we compared our results to the Breast Cancer Screening Consortium (BCSC) across a random sample of 203 studies and achieved 99.5% agreement in labeling of mammograms as positive or negative.

While this study demonstrated the feasibility and acceptable performance characteristics of a hybrid framework for processing clinical reports, there are multiple areas for future work in addition to those described above. We noticed erroneous class predictions were typically accompanied by low confidence scores. Moving forward, confidence thresholds could be implemented to flag certain cases for manual review. This would decrease the overall error rate while still dramatically reducing the amount of manual effort required for data annotation. Furthermore, alternate methods of text preprocessing such as inclusion of certain punctuation to determine end-of-sentence and numerics to capture tumor size could be considered.

Our technique's generalizability to other medical datasets also remains to be seen, particularly those that may be less structured. Further investigation on datasets of different sizes, numbers and distribution of classes, and combinations of structured and unstructured data is needed. It is also unclear whether data from multiple cross-institutional datasets can be combined without jeopardizing performance. Finally, we would like to expand this work to investigate several other automated solutions such as Microsoft Azure Text Analytics, Facebook fastText, Amazon Comprehend, and Google AutoML.

Conclusion

We demonstrate a framework capable of assigning labels to free-text pathology records using both traditional natural language processing techniques and IBM Watson. IBM Watson performed favorably for reports under 1024 characters which comprised 92% of cases in our dataset, thus significantly lowering the barrier to entry and domain knowledge required for natural language processing. Future work will focus on expanding this process to other medical records such as radiology reports and clinical notes as well as testing other automated solutions from Facebook, Google, Amazon, and Microsoft. We hope to design an automated pipeline for large-scale clinical data annotation so that existing clinical records can be efficiently utilized for development of deep learning algorithms.

Funding Information DH and MP are supported by the National Cancer Institute of the NIH under Award Number UH2CA203792, and the National Library of Medicine of the NIH under Award Number 1U01LM012675. HT was supported by NIH T32 Grant 5T32EB001631-10. AL was supported by the UCSF RAPtr fund.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, Ankerst DP, Kaklamani VG, Rodriguez AA, Wong STC, Chang JC: Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer 123(1):114–121, 2016. https://doi.org/10.1002/cncr.30245
2. Töyräs J, Kröger H, Jurvelin JS: Bone properties as estimated by mineral density, ultrasound attenuation, and velocity. Bone 25(6): 725–731, 1999. https://doi.org/10.1016/S8756-3282(99)00221-5

3. Tarver T: Cancer facts & figures 2012. American Cancer Society (ACS). J Consum Health Internet 16(3):366–367, 2012. https://doi.org/10.1080/15398285.2012.701177.
4. Bleyer A, Baines C, Miller AB: Impact of screening mammography on breast cancer mortality. Int J Cancer 138(8):2003–2012, 2016. https://doi.org/10.1002/ijc.29925
5. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, Henderson LM, Onega T, Tosteson ANA, Rauscher GH, Miglioretti DL: National Performance Benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. Radiology 283(1):49–58, 2017. https://doi.org/10.1148/radiol.2016161174
6. Kopans DB: An open letter to panels that are deciding guidelines for breast cancer screening. Breast Cancer Res Treat 151(1):19–25, 2015. https://doi.org/10.1007/s10549-015-3373-8
7. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521(7553): 436–444, 2015. https://doi.org/10.1038/nature14539
8. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen - van de Kaa C, Bult P, van Ginneken B, van der Laak J: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 6(1):26286, 2016. https://doi.org/10.1038/srep26286
9. Suk H-I, Lee SW, Shen D, Initi ADN: Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101:569–582, 2014. https://doi.org/10.1016/j.neuroimage.2014.06.077
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118, 2017. https://doi.org/10.1038/nature21056
11. Arfan M: Deep learning based computer aided diagnosis system for breast mammograms. Int J Adv Comput Sci Appl 8(7), 2017. https://doi.org/10.14569/IJACSA.2017.080738
12. Dhungel N, Carneiro G, Bradley AP: Deep structured learning for mass segmentation from mammograms. In: IEEE 2950–2954, 2015. https://doi.org/10.1109/ICIP.2015.7351343.
13. Dhungel N, Carneiro G, Bradley AP: Combining deep learning and structured prediction for segmenting masses in mammograms. In: Deep Learning and Convolutional Neural Networks for Medical Image Computing. Vol 58. Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing 225–240, 2017. https://doi.org/10.1007/978-3-319-42999-1_13.
14. Wang J, Yang Y: A context-sensitive deep learning approach for microcalcification detection in mammograms. Pattern Recogn 78: 12–22, 2018. https://doi.org/10.1016/j.patcog.2018.01.009
15. Dhungel N, Carneiro G, Bradley AP: A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med Image Anal 37:114–128, 2017. https://doi.org/10.1016/j.media.2017.01.009
16. Greenspan H, van Ginneken B, Summers RM: Deep learning in medical imaging: overview and future promise of an exciting new technique. arXiv. 35(5):1153–1159, 2016. https://doi.org/10.1109/TMI.2016.2553401
17. He K, Zhang X, Ren S, Sun J: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: IEEE 1026–1034, 2015. https://doi.org/10.1109/ICCV.2015.123
18. Nguyen H, Patrick J: Text mining in clinical domain. New York: ACM Press; 2016:549–558. https://doi.org/10.1145/2939672.2939720.
19. Rodríguez-González A: Extracting diagnostic knowledge from MedLine plus: a comparison between MetaMap and cTAKES approaches. Curr Bioinforma 12:1–11, 2017. https://doi.org/10.2174/1574893612666170727094502
20. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513, 2010. https://doi.org/10.1136/jamia.2009.001560
21. Garla V, Re, III VL, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, Justice A, Brandt C: The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 18(5):614–620, 2011. https://doi.org/10.1136/amiajnl-2011-000093
22. Pons E, Braun LMM, Hunink MGM, Kors JA: Natural language processing in radiology: a systematic review. Radiology 279(2): 329–343, 2016. https://doi.org/10.1148/radiol.16142770
23. Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B, Witten IH, Trigg L: Weka-a machine learning workbench for data mining. In: Data Mining and Knowledge Discovery Handbook. Boston: Springer US:1269–1277, 2010. https://doi.org/10.1007/978-0-387-09823-4_66
24. Holmes G, Donkin A, Witten IH: WEKA: a machine learning workbench. In: IEEE:357–361. https://doi.org/10.1109/ANZIIS.1994.396988
25. Ferrucci D, Levas A, Bagchi S, Gondek D, Mueller ET: Watson: beyond jeopardy! Artif Intell 199:93–105, 2013. https://doi.org/10.1016/j.artint.2012.06.009
26. Brown E. Watson: the jeopardy! Challenge and beyond. In: IEEE: 2–2, 2013. https://doi.org/10.1109/ICCI-CC.2013.6622216
27. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH: Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging 11(5):245–251, 2017. https://doi.org/10.1007/s10278-017-0021-3