Journal of Biomedical Informatics 69 (2017) 177–187

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automated annotation and classification of BI-RADS assessment from radiology reports

Sergio M. Castro, Eugene Tseytlin, Olga Medvedeva, Kevin Mitchell, Shyam Visweswaran, Tanja Bekhuis, Rebecca S. Jacobson *

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, The Offices at Baum, 5607 Baum Boulevard, BAUM 423, Pittsburgh, PA 15206-3701, USA

ARTICLE INFO

Article history:
Received 15 November 2016
Revised 12 April 2017
Accepted 14 April 2017
Available online 18 April 2017

Keywords:
Breast Imaging Reporting and Data System (BI-RADS)
Information extraction
Natural language processing
Imaging informatics
Machine learning

ABSTRACT

The Breast Imaging Reporting and Data System (BI-RADS) was developed to reduce variation in the descriptions of findings. Manual analysis of breast radiology report data is challenging but is necessary for clinical and healthcare quality assurance activities.

The objective of this study is to develop a natural language processing (NLP) system for automated BI-RADS categories extraction from breast radiology reports. We evaluated an existing rule-based NLP algorithm, and then we developed and evaluated our own method using a supervised machine learning approach. We divided the BI-RADS category extraction task into two specific tasks: (1) annotation of all BI-RADS category values within a report, (2) classification of the laterality of each BI-RADS category value. We used one algorithm for task 1 and evaluated three algorithms for task 2. Across all evaluations and model training, we used a total of 2159 radiology reports from 18 hospitals, from 2003 to 2015.

Performance with the existing rule-based algorithm was not satisfactory. Conditional random fields showed a high performance for task 1 with an F-1 measure of 0.95. Rules from partial decision trees (PART) algorithm showed the best performance across classes for task 2 with a weighted F-1 measure of 0.91 for BIRADS 0-6, and 0.93 for BIRADS 3-5. Classification performance by class showed that performance improved for all classes from Naïve Bayes to Support Vector Machine (SVM), and also from SVM to PART.

Our system is able to annotate and classify all BI-RADS mentions present in a single radiology report and can serve as the foundation for future studies that will leverage automated BI-RADS annotation, to provide feedback to radiologists as part of a learning health system loop.

1. Introduction

Breast cancer is the most common malignancy in women in the United States [1], and also in the world [2]. It is a major public health concern with considerable medical and economic burden. Early detection of breast cancer is associated with treatment at earlier stage and mortality reduction [3]. Clinical practice guidelines include regular screening mammography recommendations for women of average risk [4,5].

The Breast Imaging Reporting and Data System (BI-RADS) was developed by the American College of Radiology (ACR) to reduce variation in the radiologists' descriptions of findings used for diagnosis [6]. BI-RADS includes (1) a standard lexicon to describe anatomical features present in breast imaging, and (2) a classification system designed to categorize by likelihood of malignancy, independently to each breast (Table 1) [7–9].

In addition to its clinical use, the system is also used in research settings and as a healthcare quality assurance tool in mammography, ultrasound and magnetic resonance imaging [10]. Healthcare quality assurance programs in breast imaging provide feedback to radiologists regarding their ability to detect and diagnose lesions, with the goal of continuous performance improvement. Initiatives

Abbreviations: ACR, American College of Radiology; BI-RAD, Breast Image-Reporting and Data System; BROK, BI-RADS Observation Kit; CI, confidence interval; CRF, conditional random fields; cTAKES, clinical Text Analysis and Knowledge Extraction System; MALLET, MAchine Learning for LanguagE Toolkit; ML, machine learning; MQSA, Mammography Quality Standards Act; NB, Naïve Bayes; NLP, Natural Language Processing; SVMs, support vector machines; TIES, Text Information Extraction System; UIMA, Unstructured Information Management Architecture; WEKA, Waikato Environment for Knowledge Acquisition.

* Corresponding author.

E-mail addresses: sec113@pitt.edu (S.M. Castro), tseytlin@pitt.edu (E. Tseytlin), opm1@pitt.edu (O. Medvedeva), kevin.jude.mitchell@gmail.com (K. Mitchell), shv3@pitt.edu (S. Visweswaran), tanja.bekhuis@tcbinfosci.com (T. Bekhuis), rebeccaj@pitt.edu (R.S. Jacobson).

http://dx.doi.org/10.1016/j.jbi.2017.04.011
1532-0464/

Table 1
Description of BI-RADS assessment categories [7].

BI-RADS category	Final assessment meaning	Likelihood of breast cancer
0	Need additional imaging evaluation and/or prior imaging for comparison	Not applicable
1	Negative	Negligible
2	Benign finding	Negligible
3	Probably benign finding	<2%
4	Suggestive of abnormality	23–34%
5	Highly suggestive of malignancy	≥95%
6	Malignancy confirmed by biopsy	100%

designed to increase and improve feedback to radiologists are becoming more common, extending even beyond the Mammography Quality Standards Act (MQSA) [11] and hospital accreditation practices [12]. Current performance evaluations for radiologists use BI-RADS to summarize performance metrics in correlation with pathology reports [13] and/or standard of practice benchmarks [10].

Data needed for performance evaluation and quality reviews are stored within radiology and pathology systems, and typically as free text. Consequently, most provider organizations use a labor-intensive, manual procedure for correlating this data, including manual data entry for all breast imaging patients. The current process requires collection and coding of key data elements for efficient retrieval, follow-up of patients' outcomes, correlation of pathology results with radiologists' reports, computation of metrics based on the rates of patients' outcomes and rates of adherence to evidence-based guidelines. These steps must then be repeated at specific intervals to ensure adequate follow up for low-risk categories.

Manual analysis of these data can be quite challenging due to the large volume of screening examinations, reliance on manual abstraction and report generation, sequestration of the data in task specific databases and systems, limited access to clinical reports, patient mobility, off-site biopsies, and physicians' referral patterns. In addition, the required human effort imposes significant limits on the timeliness, granularity and flexibility of the data that can be provided for healthcare performance improvement and research purposes.

Various natural language processing (NLP) techniques can be employed to automatically identify and extract key expressions from radiology reports [14–17]. These techniques may be used independently, or in combination to accomplish different subtasks within a multi-step pipeline [18].

We sought to develop an NLP system for BI-RADS extraction as the first step in developing a complete system for correlation of radiology to pathology results with feedback to radiologists. Such a system would be able to automatically extract features of interest (e.g. BI-RADS assessment categories) from the radiology reports, correlate the findings with the pathologic findings (e.g. malignant, benign, high risk), and then present this information to the clinician in a meaningful way to support performance improvement. The envisioned system provides an example of a single learning health system loop. Learning health system approaches aim to continuously improve practice by capturing data at the various levels of clinical practice and efficiently use the data to change practice [19,20].

There have been only a few publications related to the extraction of BI-RADS features from free-text radiology reports. Most studies have focused on extracting standardized descriptions of anatomical findings [21–23], rather than the BI-RADS categories. Only one published study aimed to specifically extract the BI-RADS category from the radiology report [24]. These authors developed the BI-RADS Observation Kit (BROK) algorithm to extract the final assessment category. BROK uses regular-expression string matches to obtain the reports' BI-RADS category and the laterality of the breast to which it is assigned, when possible.

A published report of the algorithm by the developers showed high performance with recall of 100.0% (95% confidence interval (CI), 99.7, 100.0%) and precision of 96.6% (95% CI, 95.4, 97.5%). However, performance in this study was measured against radiology reports that were randomly sampled by imaging technique. This sampling strategy would almost certainly over-represent the more common low-risk categories (BI-RADS 1 and 2) and under-represent the less common high-risk categories (BI-RADS 3, 4, and 5), which are the categories of greatest interest for performance improvement.

We first sought to evaluate whether BROK could be used to extract all BI-RADS categories present in a radiology report. On the basis of our findings, we then chose to develop and evaluate our own information extraction method for BI-RADS categories extraction using a supervised machine learning approach.

2. Materials and methods

2.1. Data source

All radiology reports used in this study were obtained from the University of Pittsburgh Text Information Extraction System (TIES) [25,26]. TIES is a de-identified database comprising approximately 24 million radiology reports from all 18 University of Pittsburgh Medical Center (UPMC) hospitals from 2003 to 2015.

2.2. Radiology reports

We included a total of 2159 radiology reports for all evaluations and model training described in this study. Reports included screening mammograms, diagnostic mammograms, breast ultrasounds, computed tomography, and magnetic resonance imaging. To avoid potential re-use bias, we created four different datasets and used them in different stages of our research and for different purposes (Table 2). The selection of the documents was based on BROK algorithm output. Use of the BROK algorithm in the selection step was necessary in order to obtain sufficient cases of each of the BI-RADS categories.

There was no overlap among the four datasets. A total of 1560 (72.2%) reports were used for the evaluation of the rule-based approach (BROK), and a total of 599 (27.8%) were used for the development and evaluation of the machine learning approach. Section 2.3 provides a detailed description of the development and use of datasets 1, 2, and 3. Section 2.4 provides a detailed description of the development and use of dataset 4. For each dataset, we manually classified (datasets 1–3) or manually annotated (dataset 4) BI-RADS category values and laterality.

2.3. Rule based approach

We used three different sets of breast radiology reports for the evaluation of the BROK rule-based approach (Table 2). For all datasets, selected reports were manually reviewed by one author (SC) to determine a document-level BI-RADS classification with laterality (e.g., Left, Right, Bilateral, or Nonspecific) per document, congruent with the classification scheme used in the output of the BROK system. Performance metrics (precision, recall and accuracy) were determined at the document level for each individual final BI-RADS category present in a single report, and overall for all final BI-RADS categories by comparing the BROK classification against this manual classification.

Table 2
Description of datasets.

Dataset	Number of radiology reports	Description	Use
Dataset 1	480	Random sample of radiology reports stratified by BI-RADS final category (1–6) using the BROK software, and then manually classified at document level for BI-RADS final category	Evaluation of BROK (rule-based approach) baseline performance BROK error analysis BROK rule adaptations
Dataset 2	600	Random sample of radiology reports flagged as ambiguous by BROK, and then manually classified at document level for BI-RADS final category	BROK error analysis BROK rule adaptations
Dataset 3	480	Random sample of radiology reports stratified by BI-RADS final category (1–6) using the BROK algorithm, and then manually classified at document level for BI-RADS final category	Evaluation of rule-based approach performance on BI-RADS category extraction (BROK vs. adapted BROK)
Dataset 4	599	Random sample of radiology reports stratified by BI-RADS final category (1–6) using the BROK algorithm and then manually annotated at mention level following annotation guidelines	Development and evaluation of BI-RADS category value annotator, and BI-RADS laterality classifier (machine learning approach)

We evaluated the baseline performance of BROK on dataset 1, which contained a stratified sample of 80 random radiology reports per BI-RADS category from 1 to 6 (total 480 reports). We excluded reports with BI-RADS final category of 0 from the selection criteria because these patients will, by definition, undergo further imaging studies, and therefore BIRADS 0 classes can be expected in the documents containing other BIRADS classes.

We then performed an error analysis for all cases in the first dataset where the BROK algorithm and the manual classification disagreed. To evaluate potential false negatives, we also obtained a second dataset containing a random sample of 600 reports classified as ambiguous by BROK (Table 2). These included reports with the BROK defined outputs of "multiple BI-RADS without laterality found", "conflict with bilateral/overall/combined BI-RADS", "multiple (left/right/bilateral) BI-RADS found" or "no BI-RADS category found". Error analyses on both the dataset 1 and dataset 2 were used to adapt BROK algorithm's regular expressions.

In the final step, we evaluated the adapted BROK system on a third dataset (Table 2) that contained a stratified sample of 80 randomly selected radiology reports per BI-RADS category from 1 to 6 (total 480 reports) (Table 2). We compared performance of the baseline system against the adapted system.

2.4. Machine learning approach

We also developed a second annotator using a machine learning (ML) approach, which annotated and classified multiple BI-RADS mentions within a given document. We first developed an annotation guideline and used it to create a manually annotated corpus for development, training and testing the system, based on dataset 4 (Table 2). On one subset of this data, we trained our supervised ML methods to extract BI-RADS information from the radiology reports. Finally, we evaluated the model performance and examined the resulting errors on held-out data. Fig. 1 provides a summary of the methods for this aspect of the study.

2.4.1. Annotation guidelines and schema

The annotation schema and annotation guidelines were designed using a traditional, cyclic and iterative approach. In the first step of the annotation process, two expert annotators independently applied the schema to 15 different documents for annotation training, discussed issues that arose, leading to changes in the annotation guidelines and schema. This second version of the schema was applied to 30 new documents, and Inter-Annotator Agreement (IAA) was calculated using Cohen's kappa for all the instances annotated in the schema (Fig. 3).

The annotators met to review the disagreements and arrived at consensus on the changes to the schema and the guidelines. This process was performed until the annotators achieved an IAA $\geq$ 0.85. The final annotation schema included the BI-RADS category value (0–6), the BI-RADS category descriptor (value meaning), the BI-RADS laterality (Table 3) and variations of BI-RADS acronyms.

2.4.2. Dataset

Dataset 4 was used for the development of the ML-based annotator and was produced from a stratified sample of 100 randomly selected radiology reports per BI-RADS category from 1 to 6 (total 600 reports). One sampled report was excluded from the study because it was erroneously labeled as a radiology report when it fact it was a pathology report, yielding a total corpus of 599 documents. Documents were annotated using the Anafora annotation tool [27]. Anafora is a web-based annotation tool with a human-readable XML output.

We used an annotation approach based on the assumption that single annotation of additional data can increase the amount of training data without loss of system performance [28]. The annotation task was divided into three rounds (Fig. 3). Within each round, both annotators created annotations on two sets of documents. One set contained documents that were annotated by both annotators, while the other set contained documents that were annotated only by a single annotator (either Annotator 1 or Annotator 2). The double-annotated set was used to calculate the IAA at intervals throughout the annotation process to assure a high level of agreement throughout the annotation task. The final gold standard combined the annotations of both annotators. Discrepancies presented during the annotation phase were discussed after the IAA calculation, and one was selected by consensus for use in the gold set.

2.4.3. Preprocessing radiology reports

We used the clinical Text Analysis and Knowledge Extraction System (cTAKES) [29] for preprocessing steps. cTAKES is an open-source NLP system for information extraction from electronic medical records free-text. Annotations created by cTAKES were used to create relevant features for the machine learning algorithms.

We made two minor changes to existing cTAKES annotators to be used in the preprocessing steps, based on our initial evaluations. First, we modified the cTAKES sectionizer to identify custom headers that are not included in the CDA/HL7 standard. This list of headers was based on those identified by TIES within the radiology reports. Second, we adjusted the cTAKES tokenizer to distinguish when the "." character was used as a punctuation mark versus a decimal mark, when this character followed a number.

We also developed two new Unstructured Information Management Architecture (UIMA) rule-based annotators to (1) label BI-RADS anchor expressions, such as those that were manually

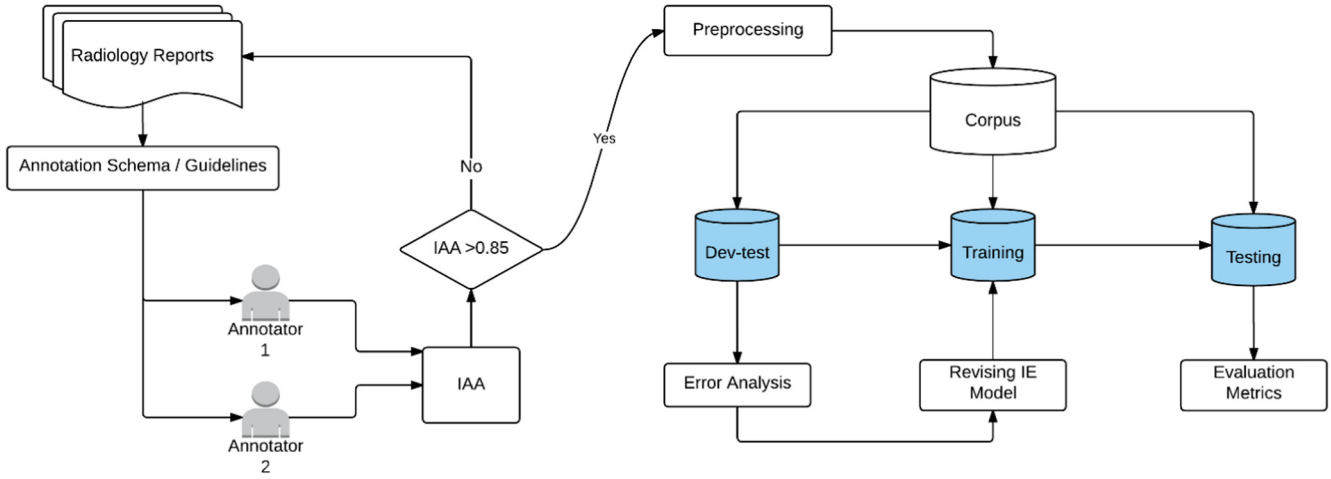

Fig. 1. Methods for development of the two-step machine learning annotator.

Table 3
Description of BI-RADS laterality.

Entity	Definition
Left BI-RADS	BI-RADS category assigned to the left breast
Right BI-RADS	BI-RADS category assigned to the right breast
Bilateral BI-RADS	BI-RADS category assigned to both breasts
Non-specific BI-RADS	Unspecified BI-RADS category, ambiguous as to whether affected breast is left or right
Overall BI-RADS	Corresponds to the most abnormal BI-RADS of the two breasts, based on the highest likelihood of malignancy. It is usually found after the detailed description of the BI-RADS category for each breast. In some cases, it is the only BI-RADS class in the report

annotated (e.g. 'BIRADS', 'BI-RADS', 'ACR code') and (2) annotate simple time expressions, e.g. '6-month', '1-year', '12:30 AM'.

The final cTAKES preprocessing pipeline consisted of the following sequential modules: (1) *sectioning* to identify the headers and annotate the various sections of the radiology report, (2) *sentence splitting*, (3) *tokenizing* to split the sentences into tokens and to classify each lexical unit into different token types, and (4) *section filtering*. In the tokenizing module, we also included the *context dependent tokenizer* to identify roman numerals that are sometimes used in the text to specify BI-RADS category values. The cTAKES *Context dependent tokenizer* creates annotations from one or more tokens, using surrounding token types as clues.

2.4.4. Model development

We divided the BI-RADS category extraction task into two specific tasks (Fig. 2). The first task was to develop an automated method for annotation of all BI-RADS risk assessment values within the breast radiology reports (including screening mammogram, diagnostic mammogram, breast ultrasound and breast MRI). The problem was formulated as a sequence prediction problem. All tokens within the text were classified as 'BI-RADS category value' or 'not BI-RADS category value', based on contextual information in the surrounding line of text. We trained a linear chain Conditional Random Fields (CRF) model to perform this task [30]. The second task was to develop an automated method for classifying the resulting annotations as Left, Right, Bilateral, Overall, or Nonspecific. The problem was formulated as a multi-class classification problem. We trained and compared multiple classifiers to perform this task.

2.4.5. Derived datasets for training and testing

We split dataset 4 as follows: a training set containing 60% of all instances, a development set containing 10% of all instances, and a test set containing 30% of all instances. The training set was used to train the classifiers, and the test set was used for the final evaluation including the final error analysis. For task 1, we used the development set for initial model building, preliminary experiments, and feature development, construction and selection. Consequently, the development set was excluded from the training and testing splits. Error analysis was performed throughout the development process, and the preprocessing system or model was subsequently adjusted. For task 2, we did not utilize the development set, and consequently we added it to the test set to increase the sample size and variance of documents used in the final evaluation.

2.4.5.1. Task 1 – Development of a BI-RADS category value annotator. We used the open-source MAchine Learning for LanguagE Toolkit (MALLET) [31] to build a linear chain CRF model for BI-RADS token annotation. CRFs are a common method used for information extraction that have been successfully applied in a wide variety of clinical text and reports [32–35]. Linear chain CRFs consider that labels of adjacent tokens are dependent on each other, therefore they are used to predict a sequence of labels given a sequence of tokens.

We used the following set of cTAKES-derived labels as features to train the model: (1) report section, (2) token type, and (3) context dependent token type. We used an iterative approach to feature construction and selection. For each set of features, we trained the model on the training set, then computed evaluation metrics and performed error analysis on the development set. We used our error analysis to identify causes of false negatives and false positives, and these insights were employed to improve the preprocessing pipeline, and to add or change features used in the subsequent model. Feature selection was expert-determined and refined using the results of the error analysis. After we reached an acceptable level of performance, the final model was tested once on the test set.

2.4.5.2. Task 2 – Development of a BI-RADS laterality classifier. We used the open-source Waikato Environment for Knowledge Acquisition (WEKA) [36] to implement three different BI-RADS laterality classifiers: Naïve Bayes (NB) [37], Rules from partial decision trees (PART) [38], and Support Vector Machine (SVM) [39]. We tested the ability of each model to classify the BI-RADS category value

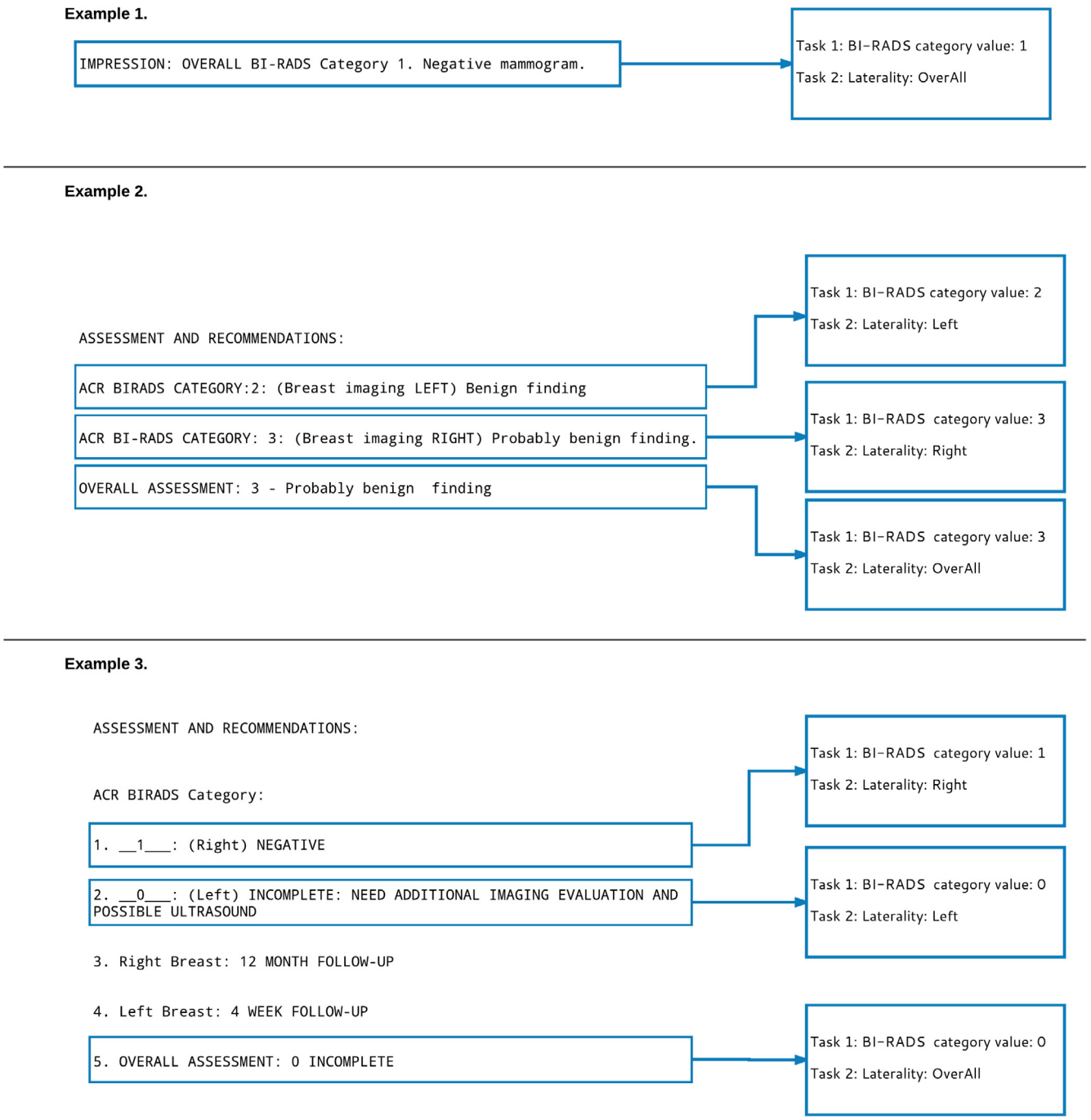

Fig. 2. Example radiology reports of increasing complexity with multiple BIRADS statements depicting the two separate tasks.

detected in the previous step as left, right, bilateral, overall, or non-specific. We chose NB because it is the simplest type of Bayesian network and makes the assumption of conditional independence of the predictors on the target variable [37,40]. PART algorithm provides an outcome that is easy to interpret. It uses partial decision trees to generate a decision list and employs a separate-and-conquer approach in each iteration, making the "best" leaf into a rule [38]. We trained the SVM classifier using sequential minimal optimization (SMO) [39,41,42] with a first-order polynomial kernel. This linear method produced the best performance when compared against a number of alternative kernels.

Each classifier was trained with an identical set of features, automatically generated as cTAKES output. These features included:

- **Bag-of-word (BoW)**. Each unique word token was treated as a feature. We included only the word tokens present in the line of the report with the BI-RADS category value as determined in Task 1. We did not normalize or remove stop words.
- **Total number of BI-RADS category value annotations.** This value was obtained by counting the number of BI-RADS category values present in the report.

- **BI-RADS category value.** We used the BI-RADS value (0–6) assigned by the radiologist and automatically extracted in Task 1.
- **BI-RADS sequence.** We assigned a rank order to each BI-RADS token annotation, based on its order of appearance in the report (e.g. first, middle, last). When there was only one BI-RADS category value in the report, it was considered to be the Last BI-RADS annotation.
- **Imaging Study type.** This was the procedure type, extracted from the report title in the radiology report system or the *Technique* description section. Possible values were *mammogram*, *ultrasound*, or *magnetic resonance imaging*.
- **Laterality.** This feature was obtained from the report title in the radiology report system or the *Technique* description section, and specified whether the study was performed on one breast or both breasts.
- **Breast(s) studied.** The anatomic location was obtained from the report title in the radiology report system or the *Technique* description section. It specified whether the study was performed on the *left breast* or *right breast*.
- **Laterality word token counts.** These were three additional features obtained by counting the mention of the word tokens "left", "right", and "bilateral" throughout the complete report.

2.4.6. Statistical analysis

We used standard metrics to evaluate the performance of both the BI-RADS annotator and classifier, including recall, precision, F1-measure, and accuracy. We separately measured performance for each BI-RADS laterality category.

For task 1 (BI-RADS category value annotation), a given annotation was considered a true positive if the model correctly predicted a token as 'BI-RADS value annotation' when compared with gold annotation, and true negative if the model correctly predicted a token as 'Not BI-RADS value' annotation, when compared with the gold annotation. False negatives were defined as cases in which the model incorrectly predicted 'Not BI-RADS value', when compared with the gold annotation. False positives were defined as cases in which the model incorrectly predicted 'BI-RADS value annotation', when compared with the gold annotation.

For task 2 (BI-RADS laterality classification), a given classification was considered a true positive if the classification (using the unordered labels 'Left', 'Right', 'Bilateral', 'Overall', 'Nonspecific') was identical to the gold annotation. True negatives were the number of correctly recognized cases that do not belong to the specific class we are trying to predict. If the classification was not identical, then the classification was an error. For this multi-class classification problem, we used micro-averaging [43] in calculating general performance metrics and classification F1 measure to compare classifiers. We measured performance for all BI-RADS category values and also for a subset of BI-RADS category values (categories 3, 4, and 5). We separately evaluated for this latter subset because these BI-RADS categories indicate an increased risk for cancer and are followed by some subsequent clinical action (re-imaging after some interval, additional imaging or biopsy). Given our anticipated use case, discriminative performance is most important in this subset.

3. Results

3.1. Rule based approach (BROK)

3.1.1. Baseline algorithm performance

Overall algorithm performance in the Dataset 1 showed that the original BROK performed relatively poorly in this sample (Table 4), with an F1 value of 0.55, favoring precision (0.69) over recall (0.46).

Table 4
BROK performance metrics.

System Dataset	Original Dataset 1	Original Dataset 3	Adapted Dataset 3
Accuracy	0.50	0.59	0.74
Recall	0.46	0.62	0.78
Precision	0.69	0.64	0.81
F-1	0.55	0.63	0.79

BROK accurately classified BI-RADS category for the left breast in 50.2% of the reports and for the right breast in 49.8% of the documents. Most of the reports contained a final BI-RADS category for each individual breast.

3.1.2. Comparison of BROK and adapted BROK

Following our initial error analysis, we then compared the performance of the original version of BROK against the adapted version of BROK (Table 4). We observed some variation in performance of the original BROK across datasets, with increased and more balanced performance in dataset 3. Adaptations we made to the algorithm had a positive effect on performance when compared to the original system.

3.1.3. BROK error analysis

The algorithm was unable to classify 131 of the BI-RADS instances present in Dataset 1, from which 74 were due to errors in the extraction of the right breast BI-RADS categories and 57 to errors in the extraction of left breast BI-RADS categories. Nearly three quarters of the extraction errors were present in documents with final category representing higher malignancy risk: BI-RADS category 4 (22.1%), BI-RADS category 3 (19.1%), BI-RADS category 5 (16%), and BI-RADS category 6 (16%).

In datasets 1 and 2, reports that contained data extraction errors shared one or more of the following features: multiple BI-RADS categories for each breast, difficulties in the preprocessing, addenda, and multiple imaging techniques in a single report. Most of data extraction errors resulted from variations in the reporting template for BI-RADS categories. We grouped the errors into

Table 5
Description of most common data extraction errors using BROK.

Error	Definition	Error source	Definition
Erroneous bilateral assignation	The algorithm erroneously assigned a category to both breasts	Removal of numerals from a list	The template allowed BI-RADS categories to be at the beginning of a sentence. BROK removed these numerals by considering it part of a list but this impedes identification
BI-RADS category was not found – regular expression trigger word not found	BROK did not assign any BI-RADS category because it failed to identify a trigger word	BI-RADS acronym not included	Some variations of the BI-RADS acronym (e.g. 'BI RADS', 'ACR code') were not included in the regular expressions
Laterality contextual errors	The tool failed to identify the correct laterality because it was not within the token distance specified in the algorithm	Distance between critical terms	The tool failed to identify the correct laterality because the term was too far from the BI-RADS trigger of the regular expression

various categories, and determined the most likely source (Table 5). On the basis of the error analysis, we adapted the BROK algorithm in three ways: (1) removed the existing rule that suppressed statements beginning with outline numbers, (2) included additional terms in the regular expressions, and (3) increased the distance between BI-RADS anchor, laterality and category. The adapted version of the algorithm, tested in dataset 3, showed continued limitations because it relied too heavily on the terms in the regular expressions, which were frequently not present within the same line.

We observed that a major limitation of BROK was that it could only assign classifications at the document level. For our future system, we anticipated the need to evaluate multiple lesions on radiology studies and correlate these lesions with the associated pathologic findings. Therefore, a basic requirement for methods development was annotation at the mention level to ensure that both breasts could be separately evaluated.

3.2. Machine learning approach

3.2.1. Annotation process and corpus development

A total of 120 documents had double annotation (annotation by two individuals) and a total of 479 had single annotation (annotation by one individual). We obtained high levels of agreement from the beginning of the process which was sustained throughout the annotation process (Fig. 3). The final corpus contained 1014 annotated BI-RADS instances in 599 radiology reports (Table 6). One report was excluded from the corpus because it turned out to be a pathology report. The final IAA across all 120 double annotated documents was 0.95.

We split the data into three sets: a training set containing 60% of all instances (608 BI-RADS instances, 368 documents), a development set containing 10% of all instances (101 BI-RADS instances, 58 documents), and a test set containing 30% of all instances (305 BI-RADS instances, 173 documents). BI-RADS instances across different categories were uniformly distributed, and corresponded to approximately 10% of all numeric tokens present in the corpus. Distributions were similar across all data sets (Table 6).

3.2.2. Evaluation of BI-RADS token annotator (Task 1)

Only three features were included for training of the baseline model to obtain the simplest model possible and improve performance over it: (1) ***section***, (e.g., *technique, findings, impression,* or *recommendation*); (2) ***token type***, (e.g., *word, punctuation, symbol, newline, contraction,* or *numeric*); and (3) ***context dependent token type***, (e.g., *roman numerals, ranges*, or *measurement*). The unit of analysis was the token. A numeric token is defined as a consecutive series of digits.

Error analysis in the development set showed that failures in the *sectionizer* altered the algorithms' ability to detect BI-RADS categories, and that time expressions (12-h format) and time-length expressions (e.g., "6-months") followed a sequence that was very similar to the BI-RADS category. We therefore adjusted the section headers to annotate only the *impression*, *recommendation*, and *addendum* sections in the section feature. We also developed two UIMA rule-based annotators, one for BI-RADS anchor words and one for time expressions. We implemented changes sequentially and observed a continuous improvement in the overall performance in the development set. The final model demonstrated a recall of 0.93, precision of 0.98 and F1 measure of 0.95. A more detailed account of the model's performance is presented in Table 7.

3.2.3. Evaluation of BI-RADS annotation classifier (Task 2)

Average classification performance was similar in the test set across all models. For all BI-RADS token annotation values (0–6), NB demonstrated a weighted F1 measure of 0.83, SVM showed weighted F1 measure of 0.89, and the PART model showed weighted F1 measure of 0.91. For the BI-RADS annotations subset (values 3–5), NB demonstrated weighted F1 measure of 0.87,

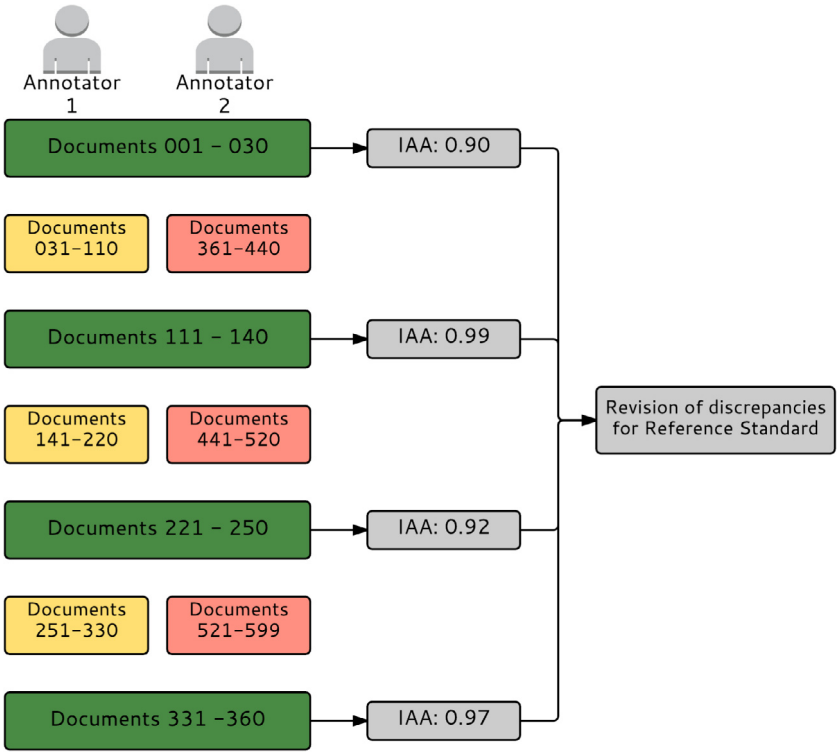

Fig. 3. Corpus annotation process and inter-annotator agreement.

Table 6
BI-RADS corpus distribution.

Corpus split	Total documents	Total number of word tokens	Total number of numeric tokens (machine annotations)	Total number of BI-RADS tokens (gold annotations)	BI-RADS category						
					0	1	2	3	4	5	6
Training	368	105,333	7588	608	83	79	129	100	90	77	50
Development	58	16,311	1189	101	12	15	16	32	14	7	5
Test	173	55,711	4077	305	48	34	67	50	46	36	24
Total	599	177,355	12,854	1014	143	128	212	182	150	120	79

Table 7
BI-RADS token annotator performance.

Task	Features	Output	TP	TN	FP	FN	Recall	Precision	F1
BI-RADS token annotator	Segment, TokenType, ContextToken, BI-RADS anchor, time expression	BI-RADS category	283	76,224	7	22	0.93	0.98	0.95

SVM showed a weighted F1 measure of 0.9, and the PART model showed a weighted F1 measure of 0.93. PART model performance was slightly superior and more balanced across the classes. Further investigation of the classification performance by class showed that performance improved for all classes from NB to SVM, and also from SVM to PART.

Bilateral BI-RADS was the class with the highest rate of classification errors for all models. Most of the errors were due to the lack of explicit indication of the class within the reports or the feature derived from the study name. Table 8 presents a detailed description of the classification performance of the different models. Although we sought to include 'non-specific' as a fifth class, the total number of instances of this class (N = 2) were too low and assigned to the training set, and therefore could not be properly evaluated in Table 8.

4. Discussion

We evaluated an existing method for the extraction of final BI-RADS assessment categories at the document level, and then developed our own multi-model NLP system for the annotation and classification of all BI-RADS assessment categories present in a single radiology report.

Table 8
BI-RADS classifiers performance.

Classifier	BI-RADS values	Output	TP	TN	FP	FN	Recall	Precision	F1
BI-RADS class Naïve Bayes	0–6	Left BI-RADS	97	279	21	9	0.92	0.82	0.87
		Right BI-RADS	81	292	15	18	0.82	0.84	0.83
		Bilateral BI-RADS	45	325	20	16	0.74	0.69	0.71
		Overall BI-RADS	113	254	12	27	0.81	0.90	0.85
		Nonspecific BI-RADS	0	404	2	0	0.00	0.00	0.00
		Weighted average							**0.83**
	3–5	Left BI-RADS	46	130	10	0	1.00	0.82	0.90
		Right BI-RADS	50	124	6	6	0.89	0.89	0.89
		Bilateral BI-RADS	3	173	4	6	0.33	0.43	0.38
		Overall BI-RADS	64	108	3	11	0.85	0.96	0.90
		Nonspecific BI-RADS	0	186	0	0	0.00	0.00	0.00
		Weighted average							**0.87**
BI-RADS class SVM	0–6	Left BI-RADS	99	278	22	7	0.93	0.82	0.87
		Right BI-RADS	90	299	8	9	0.91	0.92	0.91
		Bilateral BI-RADS	49	338	7	12	0.80	0.88	0.84
		Overall BI-RADS	124	259	7	16	0.89	0.95	0.92
		Nonspecific BI-RADS	0	406	0	0	0.00	0.00	0.00
		Weighted average							**0.89**
	3–5	Left BI-RADS	46	128	12	0	1.00	0.79	0.88
		Right BI-RADS	50	129	1	6	0.89	0.98	0.93
		Bilateral BI-RADS	6	174	3	3	0.67	0.67	0.67
		Overall BI-RADS	66	109	2	9	0.88	0.97	0.92
		Nonspecific BI-RADS	0	186	0	0	0.00	0.00	0.00
		Weighted average							**0.90**
BI-RADS class PART	0–6	Left BI-RADS	101	293	7	5	0.95	0.94	0.94
		Right BI-RADS	89	295	12	10	0.90	0.88	0.89
		Bilateral BI-RADS	51	339	6	10	0.84	0.89	0.86
		Overall BI-RADS	127	253	13	13	0.91	0.91	0.91
		Nonspecific BI-RADS	0	406	0	0	0.00	0.00	0.00
		Weighted average							**0.91**
	3–5	Left BI-RADS	45	138	2	1	0.98	0.96	0.97
		Right BI-RADS	49	128	2	7	0.88	0.96	0.92
		Bilateral BI-RADS	9	174	3	0	1.00	0.75	0.86
		Overall BI-RADS	70	105	6	5	0.93	0.92	0.93
		Nonspecific BI-RADS	0	186	0	0	0.00	0.00	0.00
		Weighted average							**0.93**

To the best of our knowledge, there are no other published reports that may be directly compared to our results. Many studies have developed tools to extract useful information from breast radiology reports using NLP techniques [21,22,44–47]. However, most of them have been limited to extraction of observed features from the reports, such as density, shape, and calcifications, and they relate exclusively to mammography reports. Our work differs because our goal was to extract the BI-RADS categories rather than the description of the findings. Only the publication of the BROK algorithm [24] had objectives similar to our research. An important difference between our goals and the goals of the BROK authors is that we specifically sought to annotate all mentions of BI-RADS categories in a report including both breasts as well as overall assessment. In contrast, the authors of BROK focused only on the final BI-RADS category for a given report. Consequently, it is not meaningful to directly compare our results to those of these previous authors.

4.1. Rule based approach (BROK)

Our evaluation of BROK showed a very different performance profile when compared to the previously published evaluation [24]. Even the adaptations we performed on the algorithm were not sufficient to obtain a satisfactory classification performance for our purposes. Two factors could have contributed to this observed difference between these evaluations.

First, the increased number of hospitals, increased time-frame and increased complexity of reports in our evaluation, when compared to the original evaluation could explain the decreased performance that we observed. In general, rule based approaches can be brittle, because the algorithm must account for many different potential patterns, which must be known a priori. We note that the regular expression approach may be improved by the implementation of specialty specific standardized lexicons. In the case of radiology, there is such a trend towards standardization of radiological language [48]. However, the use of templates and lexicons is still not the norm. Another important difference is that our evaluation was conducted on a sample stratified by BI-RADS category. Because the preponderance of breast imaging studies are screening mammography exams with negative/benign results [49–51], the sample used in the previously published BROK evaluation is likely to be heavily skewed towards simpler and more uniform language.

Second, we identified the level of classification of BROK to be problematic. BROK identifies the BI-RADS category across the entire document. However, cases with differing and actionable BI-RADS assessments of each breast are not uncommon. BROK also does not attempt to differentiate between a bilateral and an overall BI-RADS assessment. A bilateral assessment with high risk of malignancy has completely different implications for patients and clinicians, in comparison to an overall assessment with high risk of malignancy. This is particularly important in radiologic-pathologic correlation, where two specimens must be correlated in the case of a *bilateral* BI-RADS assessment, whereas only one specimen must be correlated in the case of an *overall* BI-RADS assessment.

Finally, since our goal is to develop the methods needed for a learning healthcare system feedback loop, we considered more complex reports associated with higher BI-RADS categories to be very important, even though they are less frequent in any large corpus. For all of these reasons, we elected to develop our own method for BI-RADS annotation and information extraction, using a machine learning approach.

4.2. BI-RADS category value annotator

The use of multi-step systems for clinical information extraction tasks is not new. Such systems have been successfully applied in the past for diverse tasks such as coreference resolution [52], and relationship extraction [33]. Our system for BI-RADS annotation and classification shares the same general structure: the first stage consists of a preprocessing step followed by a machine learning algorithm that focuses on recognition of the entity, and the second stage implements a different machine learning algorithm that uses the output of the previous stage and perform the classification task. However, although our system follows this same general structure, our approach differs in terms of the methods and the features we aimed to extract and classify.

We developed a linear chain-CRF model that demonstrated high recall, precision and accuracy for the annotation of all BI-RADS tokens within any section of a breast radiology report. Our BI-RADS token annotator is able to detect BI-RADS categories with a high level of granularity and accuracy, and is capable of identifying all the types of BI-RADS token categories (including overall and bilateral BI-RADS assessments) present in a breast imaging reports. Importantly, because we specifically sought to categorize the full range of categories, and to annotate all assessments (instead of producing a single annotation for the entire document), results from our study are not easily compared with any existing method, including BROK.

The pipeline approach is an important strength of our method which likely contributed to its high accuracy. Specifically, an automated section detection method was used to produce machine annotation for sections that were used as features for the CRF. Similarly, we used a modified tokenizer to create machine annotations for common numeric token types (including time), which were also features in our CRF. These features decreased the number of false negatives, as shown in our development set results. The inclusion of our UIMA annotators for time and BI-RADS anchors also helped to decrease the number of false positives.

Results from our error analysis showed that in rare situations, sectionizer failures at the beginning of a section were more likely to produce errors for the model. These problems could be easily solved in the future by standardizing the report's format. However, the overall impact of these errors on the system performance was quite small.

4.3. BI-RADS laterality classifiers

NB networks and SVM classifiers have been used in the past for classification of radiology reports [21,53–57]. NB networks have been used specifically for classification of breast radiology reports [58,59]. However, these studies have been limited to malignancy risk detection from BI-RADS lexicon features. SVM after a CRF detection has been successfully used for classification in biomedical text [33,52]. However, importantly, no study to date has tried to classify the reports into specific categories and with the level of granularity that we did in this study.

Our results demonstrate that we were able to assign BI-RADS laterality categories with moderate to very good performance with all of the classifiers. Our best classifier used the PART algorithm which demonstrated very good recall, precision and F1 measure for the multiclass classification of BI-RADS categories, once the annotation has been detected. Decision trees have the advantage of being relatively easy to interpret.

All models performed better for *left* class and *right* class when compared to *bilateral* class. There are many factors that could account for this observed classification behavior, including (1) the expert-constructed feature space may be biased towards the detection of left and right class, (2) variations in reporting across different radiology techniques that could produce missing data for key features to describe bilateral (3) the absence of specific features to describe bilateral and overall BI-RADS, (4) laterality coreference within the reports, and (5) ambiguity within the reports.

Performance variation among classifiers may be a reflection of the fact that we did not fully explore the entire feature space. Expert constructed features produced the potential for conflicting features, which would likely be resolved more simply with the decision tree and explain PART's better performance.

We included features such as imaging study and laterality counts to balance the assumption that a significant amount of information was present in the same text line of the report as the BI-RADS value. By including these features, we attempted to enhance semantic information and reduce the ambiguity that could be present in a single line of the report. In a recent study by Bozkurt and Rubin [60], a different approach to reduce ambiguity in BI-RADS reporting was assessed, and these researchers were able to identify mismatches between BI-RADS reported categories and descriptions of mammogram laterality. Detection of ambiguities could help reduce variability in reporting and improve performance of algorithms that aim to classify laterality.

4.4. Limitations

An important potential limitation of our methods was that we used the BROK software to sample by BI-RADS category during the development of the corpus used for development of our machine-learning base annotator. Although we manually classified or annotated all documents, it is likely that the most complex documents (classified by BROK as ambiguous due to multiple BI-RADS mentions) were not included in Dataset 4. A future prospective study of the software against manual abstractors will be needed to determine whether our machine learning based annotator is robust for such highly complex reports.

For the BI-RADS token annotator, the generalizability of our findings may also be limited because the detection step of our annotator strongly relies on the accurate performance of the preprocessing steps.

5. Conclusion

In conclusion, we have developed and evaluated a complete NLP system for automated BI-RADS annotation and classification, using a novel approach. BI-RADS classification with PART showed very good performance, and was consistent across all laterality classes. Our system is able to provide a detailed list of BI-RADS categories present in a single radiology report.

This work provides a solid foundation for future studies that will leverage automated BI-RADS annotation, to provide feedback to radiologists as part of a learning health system loop.

Funding sources

This research was supported by the National Cancer Institute (1U24CA184407). The first author (SC) was supported by the Fogarty Training Grant 1 D43 TW008443.

Conflicts of interest

At the time of this publication, the authors do not report any conflict of interest.

Acknowledgements

We thank Maria Bond at the University of Pittsburgh Department of Biomedical Informatics for preparation and review of the manuscript. We also thank Guergana Savova at Harvard Medical School and Computational Health Informatics Program and Harry Hochheiser at University of Pittsburgh Department of Biomedical Informatics for expert review of the manuscript.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2016, CA Cancer J. Clin. 66 (2016) 7–30.
[2] L.A. Torre, F. Bray, R.L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, CA Cancer J. Clin. 65 (2015) 87–108.
[3] D.A. Berry, K.A. Cronin, S.K. Plevritis, D.G. Fryback, L. Clarke, M. Zelen, et al., Effect of screening and adjuvant therapy on mortality from breast cancer, N. Engl. J. Med. 353 (2005) 1784–1792.
[4] U. S. Preventive Services Task Force, Screening for breast cancer: U.S. preventive services task force recommendation statement, Ann. Int. Med. 151 (2009) 716–726. W-236.
[5] K.C. Oeffinger, E.H. Fontham, R. Etzioni, et al., Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society, J. Am. Med. Assoc. 314 (2015) 1599–1614.
[6] American College of Radiology, Bi-Rads Committee, ACR BI-RADS Atlas: Breast Imaging Reporting and Data System, American College of Radiology, Reston, VA, 2013.
[7] M.M. Eberl, C.H. Fox, S.B. Edge, C.A. Carter, M.C. Mahoney, BI-RADS classification for management of abnormal mammograms, J. Am. Board Fam. Med. 19 (2006) 161–164.
[8] M.A. Lacquement, D. Mitchell, A.B. Hollingsworth, Positive predictive value of the breast imaging reporting and data system, J. Am. Coll. Surg. 189 (1999) 34–40.
[9] S.G. Orel, N. Kay, C. Reynolds, D.C. Sullivan, BI-RADS categorization as a predictor of malignancy, Radiology 211 (1999) 845–850.
[10] E.A. Sickles, C.J. D'Orsi, ACR BI-RADS® Follow-up and Outcome Monitoring, ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, American College of Radiology, Reston, VA, 2013.
[11] U. S. Food and Drug Administration, Mammography Quality Standards Act (MQSA). <http://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/Regulations/ucm110823.htm> (accessed April 22, 2015).
[12] J.R. Steele, D.M. Hovsepian, D.F. Schomer, The joint commission practice performance evaluation: a primer for radiologists, J. Am. Coll. Radiol. 7 (2010) 425–430.
[13] B.M. Geller, K. Kerlikowske, P.A. Carney, L.A. Abraham, B.C. Yankaskas, S.H. Taplin, et al., Mammography surveillance following breast cancer, Breast Cancer Res. Treat. 81 (2003) 107–115.
[14] T. Cai, A.A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K.K. Kumamaru, et al., Natural language processing technologies in radiology research and clinical applications, RadioGraphics 36 (2016) 176–191.
[15] L.T.E. Cheng, J. Zheng, G.K. Savova, B.J. Erickson, Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing, J. Digit. Imag.: Off. J. Soc. Comput. Appl. Radiol. 23 (2010) 119–132.
[16] A.-D. Pham, A. Névéol, T. Lavergne, D. Yasunaga, O. Clément, G. Meyer, et al., Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings, BMC Bioinformatics 15 (2014) 266.
[17] E. Pons, L.M.M. Braun, M.G.M. Hunink, J.A. Kors, Natural language processing in radiology: a systematic review, Radiology 279 (2016) 329–343.
[18] F. Liu, J. Chen, A. Jagannatha, H. Yu, Learning for Biomedical Information Extraction: Methodological Review of Recent Advances. <http://arxiv.org/abs/1606.07993>, 2016 (accessed June 26, 2016).
[19] C. Friedman, J. Rubin, J. Brown, M. Buntin, M. Corn, L. Etheredge, et al., Toward a science of learning systems: a research agenda for the high-functioning learning health system, J. Am. Med. Inform. Assoc. 22 (2015) 43–50.
[20] Institute of Medicine (US), L. Olsen, D. Aisner, J.M. McGinnis, The Learning Healthcare System: Workshop Summary. <http://www.ncbi.nlm.nih.gov/books/NBK92076/>, 2011 (accessed August 11, 2015).
[21] S. Bozkurt, J.A. Lipson, U. Senol, D.L. Rubin, Automatic abstraction of imaging observations with their characteristics from mammography reports, J. Am. Med. Inform. Assoc. 22 (2015) e81–e92.
[22] H. Nassif, F. Cunha, I.C. Moreira, R. Cruz-Correia, E. Sousa, D. Page, et al., Extracting BI-RADS features from Portuguese clinical texts, IEEE Int. Conf. Bioinfor. Biomed. (2012) 1–4.
[23] Y. Xu, J. Tsujii, E.I.C. Chang, Named entity recognition of follow-up and time information in 20,000 radiology reports, J. Am. Med. Inform. Assoc. 19 (2012) 792–799.
[24] D.A. Sippo, G.I. Warden, K.P. Andriole, R. Lacson, I. Ikuta, R.L. Birdwell, et al., Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing, J. Digit. Imag. 26 (2103) 989–994.
[25] R.S. Crowley, M. Castine, K. Mitchell, G. Chavan, T. McSherry, M. Feldman, caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research, J. Am. Med. Infor. Assoc. 17 (2010) 253–264.
[26] The TIES Cancer Research Network (TCRN) Homepage.
[27] W.-T. Chen, W. Styler, Anafora: a web-based general purpose annotation tool, in: K. Vanderwende, H. Daumé III, K. Kirchhoff (Eds.), 2013 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, The Association for Computational Linguistics, Atlanta, GA, pp. 14–19.
[28] D. Dligach, M. Palmer, Reducing the need for double annotation, in: LAW V '11 Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Stroudsburg, PA, 2011, pp. 65–73.
[29] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[30] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.
[31] A.K. McCallum, MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>, 2002 (accessed).
[32] M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, et al., A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Inform. Assoc. 18 (2011) 601–606.
[33] J. Patrick, M. Li, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, J. Am. Med. Inform. Assoc. 17 (2010) 524–527.
[34] M. Torii, K. Wagholikar, H. Liu, Using machine learning for concept extraction on clinical documents from multiple data sources, J. Am. Med. Inform. Assoc. 17 (2011) 524–527.
[35] O. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (2010) 514–518.
[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11 (2009) 10–18.
[37] G. Meyfroidt, F. Güiza, J. Ramon, M. Bruynooghe, Machine learning techniques to examine large patient databases, Best Pract. Res. Clin. Anaesthesiol. 23 (2009) 127–143.
[38] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, in: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 1998, pp. 144–151.
[39] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: S. Bernhard, Ikopf, J.C.B. Christopher, J.S. Alexander (Eds.), Advances in Kernel Methods, MIT Press, 1999, pp. 185–208.
[40] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 338–345.
[41] T. Hastie, R. Tibshirani, Classification by pairwise coupling, Ann. Stat. 26 (1998) 451–471.
[42] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (2001) 637–649.
[43] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inform. Process. Manage. 45 (2009) 427–437.
[44] H. Gao, E.J. Aiello Bowles, D. Carrell, D.S.M. Buist, Using natural language processing to extract mammographic findings, J. Biomed. Inform. 54 (2015) 77–84.
[45] N.L. Jain, C. Friedman, Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports, American Medical Informatics Association Annual Fall Symposium: American Medical Informatics Society, 1997, pp. 829–833.
[46] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, D. Page, Information extraction for clinical data mining: a mammography case study, in: IEEE International Conference on Data Mining Workshops, 2009, pp. 37–42.
[47] B. Percha, H. Nassif, J. Lipson, E. Burnside, D. Rubin, Automatic classification of mammography reports by BI-RADS breast tissue composition class, J. Am. Med. Inform. Assoc. 19 (2012) 913–916.
[48] RadLex. <http://www.rsna.org/RadLex.aspx>, 2016 (accessed June 24, 2016).
[49] H.J. Akande, B.B. Olafimihan, O.I. Oyinloye, A five year audit of mammography in a tertiary hospital, North Central Nigeria, Niger. Med. J. J. Niger. Med. Assoc. 56 (2015) 213–217.
[50] G.M. Badan, D. Roveda Júnior, C.A.P. Ferreira, O.A. de Noronha Junior, Complete internal audit of a mammography service in a reference institution for breast imaging, Radiol. Bras. 47 (2014) 74–78.
[51] S.P. Poplack, A.N. Tosteson, M.R. Grove, W.A. Wells, P.A. Carney, Mammography in 53,803 women from the New Hampshire mammography network, Radiology 217 (2000) 832–840.
[52] S.R. Jonnalagadda, D. Li, S. Sohn, S.T.-I. Wu, K. Wagholikar, M. Torii, et al., Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules, J. Am. Med. Inform. Assoc. 19 (2012) 867–874.
[53] M. Benndorf, E. Kotter, M. Langer, C. Herda, Y. Wu, E.S. Burnside, Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American College of Radiology (ACR) BI-RADS lexicon, Eur. Radiol. 2015 (2015) 1768–1775.
[54] G. Bouzghar, B.J. Levenback, L.R. Sultan, S.S. Venkatesh, A. Cwanger, E.F. Conant, et al., Bayesian probability of malignancy with BI-RADS sonographic features, J. Ultrasound Med. 33 (2014) 641–648.
[55] B. Percha, Machine learning approaches to automatic BI-RADS classification of mammography reports, in: Department of Biomedical Informatics SU, (Ed.), Standford University, 2010.
[56] G. Zuccon, A.S. Wagholikar, A.N. Nguyen, L. Butt, K. Chu, S. Martin, et al., Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the SNOMED CT ontology, AMIA Jt. Summits Transl. Sci. Proc. 2013 (2013) 300–304.
[57] S. Bozkurt, F. Gimenez, E.S. Burnside, K.H. Gulkesen, D.L. Rubin, Using automatically extracted information from mammography reports for decision-support, J. Biomed. Infor. 62 (2016) 224–231.
[58] P. Ferreira, N.A. Fonseca, I. Dutra, R. Woods, E. Burnside, Predicting malignancy from mammography findings and surgical biopsies, Proc. (IEEE Int. Conf. Bioinfor. Biomed.) 2011 (2011), http://dx.doi.org/10.1109/BIBM.2011.71.
[59] E.A. Fischer, J.Y. Lo, M.K. Markey, Bayesian networks of BI-RADS descriptors for breast lesion classification, in: Engineering in Medicine and Biology Society, 2004 IEMBS '04 26th Annual International Conference of the IEEE2004, pp. 3031–3034.
[60] S. Bozkurt, D. Rubin, Automated detection of ambiguity in BI-RADS assessment categories in mammography reports, Stud. Health Technol. Inform. 197 (2014) 35–39.