Journal of Biomedical Informatics 92 (2019) 103137

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automatic inference of BI-RADS final assessment categories from narrative mammography report findings

Imon Banerjee[a,*], Selen Bozkurt[a,b], Emel Alkim[a], Hersh Sagreiya[c], Allison W. Kurian[d], Daniel L. Rubin[a,c]

[a] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
[b] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Akdeniz University, Antalya 07059, Turkey
[c] Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA
[d] Medicine (Oncology) and Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

ARTICLE INFO

Keywords:
BI-RADS classification
Deep learning
Mammography report
NLP
Distributional semantics
Text mining

ABSTRACT

We propose an efficient natural language processing approach for inferring the BI-RADS final assessment categories by analyzing only the mammogram findings reported by the mammographer in narrative form.

The proposed hybrid method integrates semantic term embedding with distributional semantics, producing a context-aware vector representation of unstructured mammography reports. A large corpus of unannotated mammography reports (300,000) was used to learn the context of the key-terms using a distributional semantics approach, and the trained model was applied to generate context-aware vector representations of the reports annotated with BI-RADS category (22,091). The vectorized reports were utilized to train a supervised classifier to derive the BI-RADS assessment class.

Even though the majority of the proposed embedding pipeline is unsupervised, the classifier was able to recognize substantial semantic information for deriving the BI-RADS categorization not only on a holdout internal testset and also on an external validation set (1900 reports). Our proposed method outperforms a recently published domain-specific rule-based system and could be relevant for evaluating concordance between radiologists. With minimal requirement for task specific customization, the proposed method can be easily transferable to a different domain to support large scale text mining or derivation of patient phenotype.

1. Introduction

Breast Imaging Reporting and Data System (BI-RADS) was developed by The American College of Radiology in an effort to standardize mammography reporting language and assessment of the findings [1]. Yet, in clinical practice, high inter-observer disagreement (kappa value of 0.37) is reported regarding the final assessment of BI-RADS categories [2]. For instance, earlier studies have found disagreements between observers in clinically significant management (biopsy versus follow-up) in 32% of screening interpretations and in 45% after diagnostic evaluation [3]. As a consequence, several breast cancer treatment planning studies [4,5] reported significant delays in assessing mammography exam findings, which further hampered the timely management of patients and consequently caused higher mortality rate.

Computerized inference of BI-RADS category may play a key role in progressing standardized treatment planning by providing feedback to the radiologists as they create their report to help to minimize potential harms associated with variable categorization, and it can be used to accurately prioritize the follow-up based on patient phenotype. On the other hand, it may also facilitate AI-based healthcare research by offering a large scale text mining and data gathering opportunity, ultimately supporting the development of an efficient predictive model for breast cancer. However, the main challenge for a machine to assess the content of the mammography report is the lack of both the standardization and the machine interpretability of the reports written in narrative form.

Several NLP-based methods have previously been successfully applied to radiology reports, also specifically to mammography reports, for extracting information or performing automatic classification of the reports [6–10]. In Table 1, we present a comparison of the relevant works. So far, only three proposed works [11–13] have focused on classification or annotation of the reports based on BI-RADS final assessment categories. However, the main limitation of the earlier studies is the requirement for a huge amount of manual labour for creating

* Corresponding author.
E-mail address: imonb@stanford.edu (I. Banerjee).

https://doi.org/10.1016/j.jbi.2019.103137
Received 20 February 2018; Received in revised form 2 October 2018; Accepted 15 February 2019
Available online 23 February 2019
1532-0464/

Table 1
State-of-the-art NLP systems focusing on BI-RADS.

Study year	Aim	Method	Performance	Limitation
Percha, 2012 [7]	To classify mammography reports into BI-RADS breast tissue composition categories	Rule-based	> 99% classification accuracy	Only focus on breast tissue composition
Hussam, 2013 [16]	Present a BI-RADS features extraction algorithm for clinical data mining	Rule-based	97.7% precision, 95.5% recall	Does not classify BI-RADS final assessment categories
Sippo, 2013 [12]	To extract BI-RADS final assessment categories from the reports	Rule-based	Precision = 99.6% Recall = 100%	If BI-RADS category is not reported, it cannot classify or predict the BI-RADS category
Bozkurt, 2016 [6]	Using automatically extracted information from mammography reports for decision-support	Rule-based + supervised ML	Accuracy = 97.58%	Generalizability issues since the training set is from a single center and semi-structured
Castro, 2017 [11]	To extract BI-RADS final assessment categories from the reports	Rule-based + supervised ML	Precision = 98% Recall = 93%	Can only extract the BI-RADS category if BI-RADS category is reported by the mammographer
Gupta, 2018 [15]	To extract relations in an unsupervised way from radiology reports	Rule-based + unsupervised clustering	Precision = 95% Recall = 94%	Do not classify BI-RADS final assessment categories

feature extraction rules and the lack of generalizability of their approaches to multi-institutional data. In addition, they [11,12] mostly need BI-RADS score reported inside the text for successful assessment. In contrast with the previous approaches, recent advances in NLP techniques can be leveraged for the automatic classification of mammography reports by exploiting distributional semantics [14,15].

Distributional semantics methods such as word2vec [14] automatically learn the vector representation of words and phrases by capturing the context of the term in the corpus. While word embeddings have been shown to help in various NLP tasks, to the best of our knowledge, they have not been used to support the classification of mammography reports in terms of BI-RADS final assessment categories. One of the biggest challenges with employing word2vec for radiology report parsing is how to handle unknown or out-of-vocabulary (OOV) words and morphologically similar words (abbreviations, acronyms, telegraphic phases). This problem stems from the fact that synonyms and related words are used widely in radiology depending on the preferred style of radiologist, and particular words may have been used infrequently in a large radiology report corpus, even if the corpus is collected from the same institution.

We propose a semi-supervised NLP pipeline for mammography reports that combines distributional semantics with the semantic dictionary mapping technique for creating a context-aware vector representation of the free-text reports. The vector representations are utilized in a machine-learning setting to classify mammography reports based on the six BI-RADS final assessment categories. The application of semantic dictionary mapping as the basis of report parsing helps to handle OOV words, and consideration of the key-term context-window increases feature extraction efficiency, effectively targeting the complex BI-RADS phenotyping task. We compare the performance of the proposed method against a recently published rule-based information extraction system for mammography reports which focuses on classifying BI-RADS final assessment categories based on imaging observations reported in the findings section [6]. We also compared our model performance with the out-of-the-box Word2vec (with the mean of word embeddings) and showed that it added value via context-aware embedding for BI-RADS inference.

The organization of the rest of the paper is as follows. Section 2 describes the three different datasets of mammography reports used in our study. Section 3 includes a detailed description of our method and presents the most interesting aspects of all processing phases. Results are summarized and compared with recently published rule-based and Word2vec methods in Section 4, while Section 5 contains a discussion of the results. Section 6 concludes the paper.

2. Dataset

2.1. Description of the Corpus

In order to perform the study and validate the results, we collected the three non-overlapping corpora of mammography reports.

Corpus 1. Annotated reports from radTF. With the approval from the Institutional Review Board (IRB), a total of 41,142 mammography reports were extracted from the radTF database at Stanford University [17]. The radTF database was designed by two experienced radiologists and serves as an educational tool for the training of radiology residents. Of these reports, 38,665 are diagnostic mammograms (not readings of mammograms from outside facilities, descriptions of ultrasound-guided biopsy procedures, or analyses of extracted tissue specimens) and 22,109 of them contained BI-RADS codes (older reports frequently do not have BI-RADS codes). In the current study, we include the 22,109 reports annotated with BI-RADS codes, which are mainly unilateral (single-breast) diagnostic mammography reports. As the ground truth label for training and validation, we used the BI-RADS codes that were assigned by the radiologist who read the mammogram.

Corpus 2. Unannotated reports. Additionally, we assembled a

corpus of 300,000 mammography reports from our healthcare institution's clinical data repository, which contains a variety of breast abnormalities. We used this large corpus of 300,000 un-curated reports to learn the context of key-terms using an unsupervised distributional semantic approach which does not need hand-labeled data for training.

Corpus 3. Annotated report from OncoshareDB. In order to formulate an external validation set, we collected 1900 de-identified screening and mammogram reports from the Oncoshare database (OncoshareDB) [18] (345 diagnostic, 1555 screening). Oncoshare is a collaborative research database that is composed of retrospective electronic data on women treated for breast cancer in the academic medical center of Stanford University, the research institute of the community-based Palo Alto Medical Foundation (PAMF), and the research group at the regional Cancer Prevention Institute of California (CPIC). We selected only the unilateral mammogram reports which followed the BI-RADS standardization for reporting findings. Similar to the corpus 1, we extracted the BI-RADS codes that were assigned by the radiologist who read the mammogram, as ground truth labels.

2.2. Experimental setup

In Corpus 1, 22,109 reports are annotated with BI-RADS final assessment categories, which are summary codes that indicate the radiologist's level of suspicion that a malignancy is present. We have the following 7 distinct BI-RADS classes:

1. **BI-RADS 0** – Need additional imaging evaluation and/or prior mammograms for comparison;
2. **BI-RADS 1** – Negative for abnormality;
3. **BI-RADS 2** – Benign;
4. **BI-RADS 3** – Probably benign;
5. **BI-RADS 4** – Suspicious abnormality;
6. **BI-RADS 5** – Highly suggestive of malignancy;
7. **BI-RADS 6** – Known biopsy proven malignancy.

In our experiment, the reports with the BI-RADS 5 category are dropped due to an insufficient number of reports (<5%) for training a machine learning model. From the remaining reports in Corpus 1, we randomly selected 20% as the testing dataset and the remaining data for training. The distribution of BI-RADS classes within the 17,672 annotated reports in the **training dataset** and the 4419 annotated reports in the **testing dataset** is summarized as a pie-chart in Fig. 1a.

Sampling of the training dataset: To manage the class imbalance issue in the training dataset, we performed over-sampling using the Synthetic Minority Over-sampling Technique (SMOTE) [19] for under-represented BI-RADS classes 3, 4 and 6, and we performed cleaning using Edited Nearest Neighbors (ENN) [20] for over-represented BI-RADS classes 0, 1, and 2. After sampling, the size of the training dataset increased about 3%. We report the performance of the systems both with and without sampling in the Results section (Section 4).

Class-wise distribution of the external validation set: In Fig. 1b, we represent the distribution of BI-RADS classes in our external validation set, i.e. corpus 3, as a pie-chart. The BI-RADS categories were extracted by parsing the 'Impression' section of the mammogram reports where the original reader recorded the BI-RADS score while, during our validation, we used only the text from the 'Findings' section of the reports. Among the 345 reports, 61% of reports were BI-RADS 2, 26% were BI-RADS 1, and 6% were BI-RADS 6. The sample distribution for the Oncoshare mammograms does not exactly match with the distribution in the radTF dataset since the majority class in Oncoshare is BI-RADS 2, while in radTF the majority class is BI-RADS 1. Moreover, BI-RADS 6 (biopsy proven malignancy) represents 6% of the Oncoshare data.

3. The proposed pipeline

Fig. 2 presents the core processing blocks of the proposed report categorization pipeline. The pipeline produces a context-aware dense vector embedding of the whole mammography report in which two complimentary phases are combined – (i) semantic key term mapping, and (ii) context analysis using word2vec.

Following the common pre-processing steps, semantic-dictionary mapping with domain specific key-terms is used as the basis of the word vector creation process. The semantic dictionary is also used to create a context-aware vector representation of whole reports based on five span windowing of the domain-specific key-terms. Finally, a supervised classification model is trained to learn the mapping between the report vectors of the training set and ground truth labels for predicting the annotation of test cases. The majority of the pipeline is unsupervised, and only the classification block needs manually annotated BI-RADS labels. The prototype was implemented using the Python programming language and the Gensim 2.1.0 library [21]. In the following subsection, we illustrate the functionality of the core processing blocks.

3.1. Pre-processing

In the data pre-processing step, all the textual content of the mammograpy reports (Corpus 1, 2 and 3) is stemmed and converted into lower case by using the NLTK library [22]. In addition, all the stopwords, punctuation characters, words with low frequency (<50), and words with less than 2 letters are removed. The integer and floating point numbers are converted to the corresponding string representation.

In order to preserve the local dependencies, bigram collocations of all possible word-pairs are calculated for the entire pre-processed corpus based on Pointwise Mutual Information [23]. The bigrams with less than 50 occurrence are discarded and the top 1000 bigram collocations are concatenated as a single word to improve the accuracy of the word embeddings. A few examples of the resultant bigrams are: 'chest_wall', 'reduced_dose', 'weight_loss', 'weight_gain', 'normal_followup', 'trabecular_thicken', 'suspici_microcalcif', 'cell_carcinoma'.

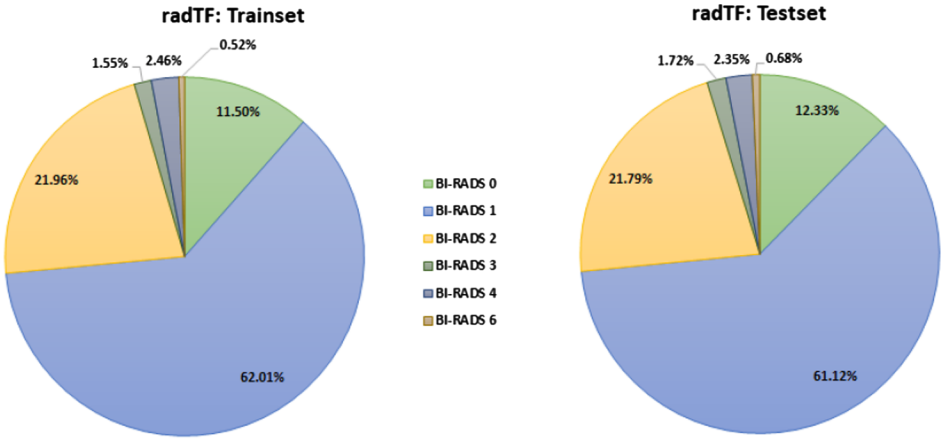
3.2. Report splitter

The BI-RADS final assessment categories are related to imaging findings, yet they are often only reported in the impression section of mammography reports. In order to evaluate the proposed report categorization approach without explicit mention of the BI-RADS category, from Corpus 1 and 2 we extract only the findings section from the mammography reports, which includes the imaging characteristics of the abnormalities. However, the impression section is excluded: thus, it does not contain any clear-cut definition of the final BI-RADS assessment class. We developed a Python-based section segmentation algorithm, Report Splitter, to separate the clinical history, findings, and impression sections. The algorithm is designed to recognize section headings and uses regular expressions to segment the reports into proper sections.

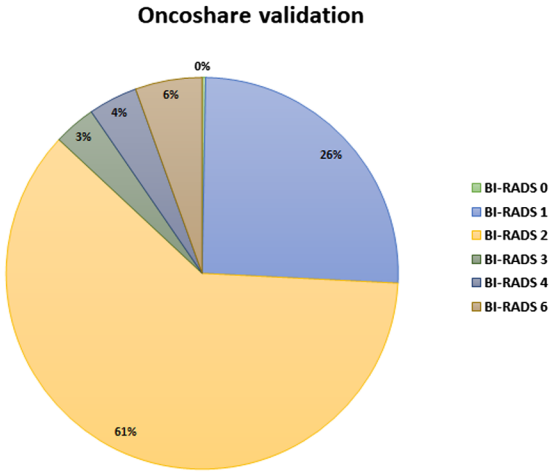
3.3. Semantic dictionary mapping

After text pre-processing and report splitting, we exploit domain ontologies to reduce term ambiguity and improve the semantic accuracy of the reports. This is done by using a lexical scanner that recognizes corpus terms which share a common root or stem with predefined terminology, and we map them to controlled terms (key-terms).

We created the domain ontology using the SPARQL (The Simple Protocol and RDF Query Language) API. We developed a SPARQL query-engine that remotely queries the RadLex lexicon hosted in the NCBO BioPortal [24] to find the key-terms provided by the domain-experts and programmatically extract a sub-tree from the RadLex



(a)



(b)

Fig. 1. Distribution of BI-RADS classes: (a) for the training and test sets in corpus 1, (b) for the external validation set in corpus 3.

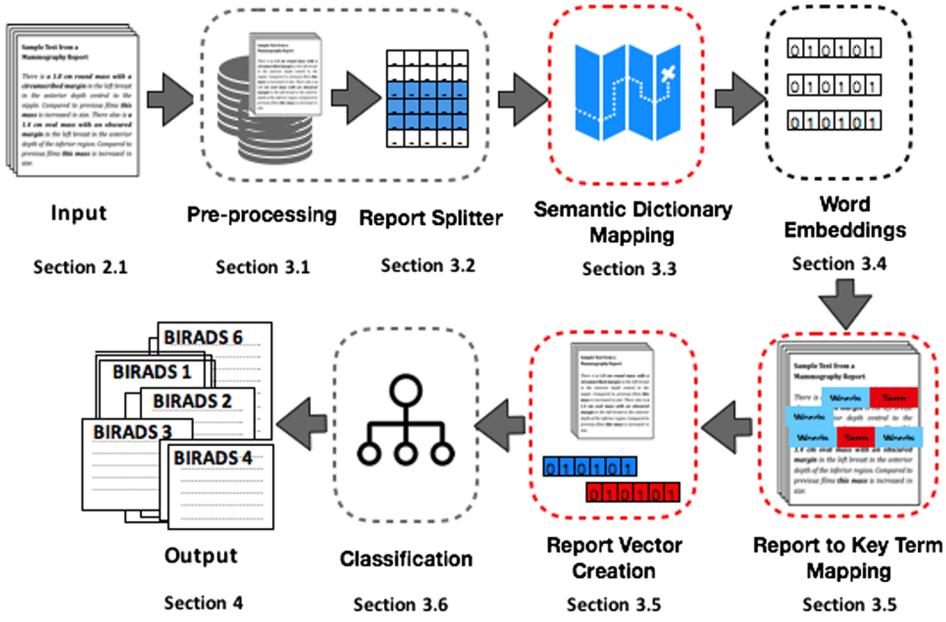


Fig. 2. The Proposed Pipeline – only the 'Classification' block depends on human annotation.

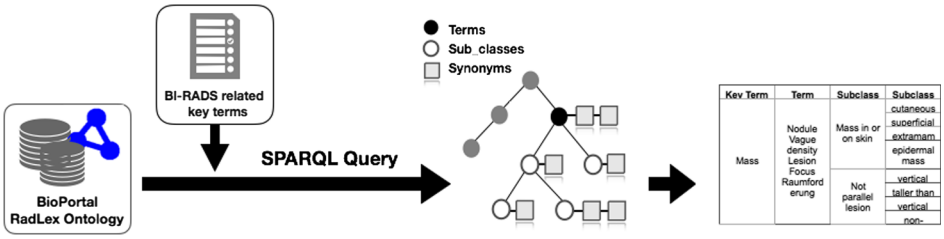


Fig. 3. Automatic domain-specific semantic dictionary generation process.

Table 2
An example part of the dictionary for the term "mass."

Key Term	Term Synonym	Subclass	Subclass Synonym
Mass	Nodule Vague density Lesion Focus Raumforderung	Mass in or on skin	Cutaneous mass Superficial mass Extramammary mass Epidermal mass
		Non-parallel mass	Vertical lesion Taller than wide mass Vertical mass Non-parallel mass

lexicon [25] that includes all the BI-RADS related terms, their synonyms, and sub-classes. The query performs pattern matching on the available graph of RadLex terminology (46,340 concepts and associated terms) and constructs a domain-specific dictionary. The dictionary is reviewed by an expert to resolve redundancy. The domain-specific semantic dictionary creation process is illustrated in Fig. 3, and an example part of the dictionary for the term "mass" can be seen in Table 2.

In addition to RadLex, we use a general publicly available terminology, CLEVER, which is designed to detect broadly applicable clinical contexts and map them to root terms, including negation (e.g., "no evidence of [condition]"), risk (e.g., "risks include [condition]"), and family related terms (e.g., "mother passes from [condition]") [26].

After combining the domain-specific key-terms and general terms derived from CLEVER terminology, we compile on a total of 325 key-terms. The key-terms are mainly used to serve two purposes in the pipeline: (i) reduce the variations in the reports via mapping, and (ii) help to generate context-aware vector representations to support report categorization (see Section 3.5).

3.4. Word-embeddings

The pre-processed mammography reports from the training set of Corpus 1 and all 300,000 reports from Corpus 2 were used to create vector embeddings for words in a completely unsupervised manner using the word2vec model [27]. The word2vec model adopts distributional semantics to learn dense vector representations of all words in the pre-processed corpus by analyzing their context. In other words, the vectors produced represent each word or phrase as a mathematical combination of the words and phrases surrounding it within a linear context window.

The semantic dictionary mapping step (Section 3.3) not only considerably reduced the size of our vocabulary by mapping the words in the corpus to key terms, but it also decreased the probability of OOV word encounters. Therefore, it facilitates the application of word2vec to directly parse radiology reports. The idea behind this is that the context of key-terms (derived from the domain-specific dictionary) should capture their true semantics and can facilitate information extraction from mammography reports. For word2vec training, we used the skip-gram model with a vector length of 300, a window width of 10, and default settings for all other parameters. No vectors were built for terms occurring fewer than 5 times in the corpus.

3.5. Context-aware vector creation

The vector creation process for each report belonging to Corpus 1 is illustrated in Fig. 4. In this step, we use the 325 key-terms defined by the BI-RADS dictionary and CLEVER terminology to identify the window-of-relevant-words for generating a context-aware vector representation of whole reports. We searched the key-terms in each report and, if a match was found, we defined its context as the term and its surrounding 4 words. We selected a small context-window (4 words) to capture the legitimate frame of the key-term appearance in the report. The choice of window size is also conditioned on the fact that the average sentence length in the mammography report corpus is 5–6 words. The context's vector was then computed as the average of its five constituent word vectors. We averaged the vectors of each word created through the trained word2vec model in the word embedding phase (see Fig. 4).

Each report vector is computed as follows: $V_{report} = \frac{1}{\| N \|} \sum_{c \in keyterms} (\frac{1}{\| n \|} \sum_{w \in context} v_w)$, where V_{report} is the report vector, v_w refers to the vector of word w inferred from the word2vec

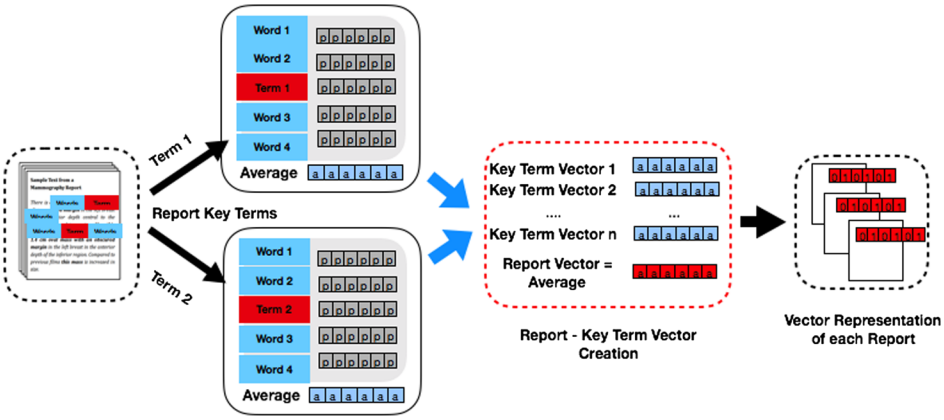


Fig. 4. Context-aware report vector creation process.

Table 3
Classification results for 4419 test reports without sampling of the training set – proposed method against a rule-based BN model and an out-of-the-box word2vec.

	Rule-based BN			out-of-the-box Word2Vec			Proposed method		
	Precision	*Recall*	*f1-score*	*Precision*	*Recall*	*f1-score*	*Precision*	*Recall*	*F1-score*
Class 0	0.78	0.47	0.59	0.94	0.59	0.72	0.91	0.92	0.91
Class 1	0.77	0.97	0.86	0.83	0.91	0.87	0.92	0.93	0.92
Class 2	0.78	0.47	0.59	0.76	0.30	0.43	0.75	0.76	0.75
Class 3	0.40	0.03	0.05	0.07	0.20	0.10	0.42	0.24	0.30
Class 4	0.50	0.41	0.45	0.34	0.61	0.43	0.67	0.75	0.71
Class 6	0.29	0.07	0.11	0.09	0.90	0.17	0.71	0.43	0.53
Avg	0.76	0.77	0.74	0.80	0.72	0.73	0.87	0.87	0.87

Table 4
Classification results for 4419 test reports with sampling of the training set – proposed method against a rule-based BN model and an out-of-the-box word2vec.

	Rule-based BN			Out-of-the-box Word2Vec			Proposed method		
	Precision	*Recall*	*f1-score*	*Precision*	*Recall*	*f1-score*	*Precision*	*Recall*	*F1-score*
Class 0	0.79	0.44	0.56	0.95	0.59	0.73	0.89	0.91	0.90
Class 1	0.75	0.74	0.75	0.84	0.91	0.87	0.97	0.82	0.89
Class 2	0.85	0.35	0.49	0.73	0.36	0.48	0.66	0.76	0.71
Class 3	0.01	0.14	0.02	0.07	0.21	0.10	0.15	0.59	0.23
Class 4	0.47	0.44	0.46	0.34	0.60	0.44	0.50	0.78	0.61
Class 6	0.06	0.23	0.09	0.10	0.90	0.18	0.42	0.79	0.55
Avg	0.75	0.59	0.64	0.80	0.73	0.74	0.81	0.81	0.83

model, n is the context window size (i.e. 5 in this study), and N is the number of key-terms present in the report.

According to Kenter et al. [28], averaging the embeddings of words in a sentence has proven to be a successful and efficient way of obtaining sentence embedding. On average, each report in the radTF dataset contains 4–5 key-terms. If a report included more than one key term, the report's individual context vector was the average of the all the key-term context vectors. We never encountered a report which did not include any of the key-terms. The most frequent appearing key-term was "breast".

3.6. Classification

The main advantage of the proposed context-aware vector representation is that it can preserve the relevant information about radiological findings reported in mammography reports, while having relatively low dimensionality. Additionally, the compact numeric representation of the free-text information allows better machine learning treatment than straight one-hot encoding of words or traditional Bag-of-words variations.

Our goal is to build a supervised classifier using the embedding of the radiological findings to automatically recognize the BI-RADS categorization of the reports. We trained a standard non-parametric Logistic Regression classifier in its default configurations (stochastic average gradient solver, intercept scaling = 1, l_2 penalty) for predicting six different BI-RADS assessment classes. We used a logistic regression classifier, though we expect that similar results could have been obtained using other supervised methods, e.g., random forests and support vector machine. The performance of the models has been measured in terms of Accuracy, Precision, Recall, and F1 score. The F1 score is calculated as the harmonic average of the precision and recall.

4. Results

As a baseline, we compared the performance of the proposed model with a recently published decision support system for breast cancer (NLP-DDS) [6,13] and an out-of-the-box Word2Vec (with an averaging of word embeddings) [14]. The NLP-DDS depends on the features extracted from the mammography reports with a set of hard NLP rules [6] and uses a Bayesian Network (BN) to predict BI-RADS category. We applied the same set of rules to extract features from our training dataset that was used to train the proposed model, in order to train a BN model from scratch.

For the out-of-the-box Word2Vec model, we trained the word2vec embedding model on the same training dataset as used by our proposed embedding method (see Section 3.4) and used the same classifier model for BI-RADS categorization – a non-parametric Logistic Regression classifier in its default configuration. We also report the performance of the rule-based, out-of-the-box Word2Vec and the proposed model on the same test dataset to be able to make a fair comparison (for the training and test set see Section 2.2). In order to derive a head-to-head comparison with the proposed method, we also validated the performance of the rule-based and out-of-the-box Word2Vec system after sampling the training data points. The performance of the models has been measured in terms of Precision, Recall, and F1 score.

4.1. Internal validation: on the radTF dataset

Tables 3 and 4 summarize side-by-side the precision, recall and F1-score of the proposed, out-of-the-box Word2Vec model, and the rule-based method, trained in a supervised manner both with and without sampling the training dataset. The performance is measured for the same test dataset. In both cases (with and without training on the sampled dataset), the proposed model out-performed both the rule-based method and the out-of-the-box Word2Vec model. Without sampling, the rule-based method achieved an average F1-score of 0.74, the out-of-the-box Word2Vec model achieved an average F1-score of 0.73, while our model scored 0.87. With sampling, average F1-scores of the rule-based method, the out-of-the-box Word2Vec model, and our model are 0.64, 0.74, and 0.83, respectively. Looking at the models' performance at the individual BI-RADS class level, the rule-based method mostly failed to classify the intermediate and malignant BI-RADS classes (BI-RADS 3, 4, and 6), even with sampled training data with equal class distribution. This may be due to the incomplete extraction of malignant features by the rule-based system, which influences the classifier to treat the malignancy as a "rare event" and ignore it. In contrast, the proposed model's performance is superior for malignant BI-RADS classes in both settings, which shows the fact that the classifier

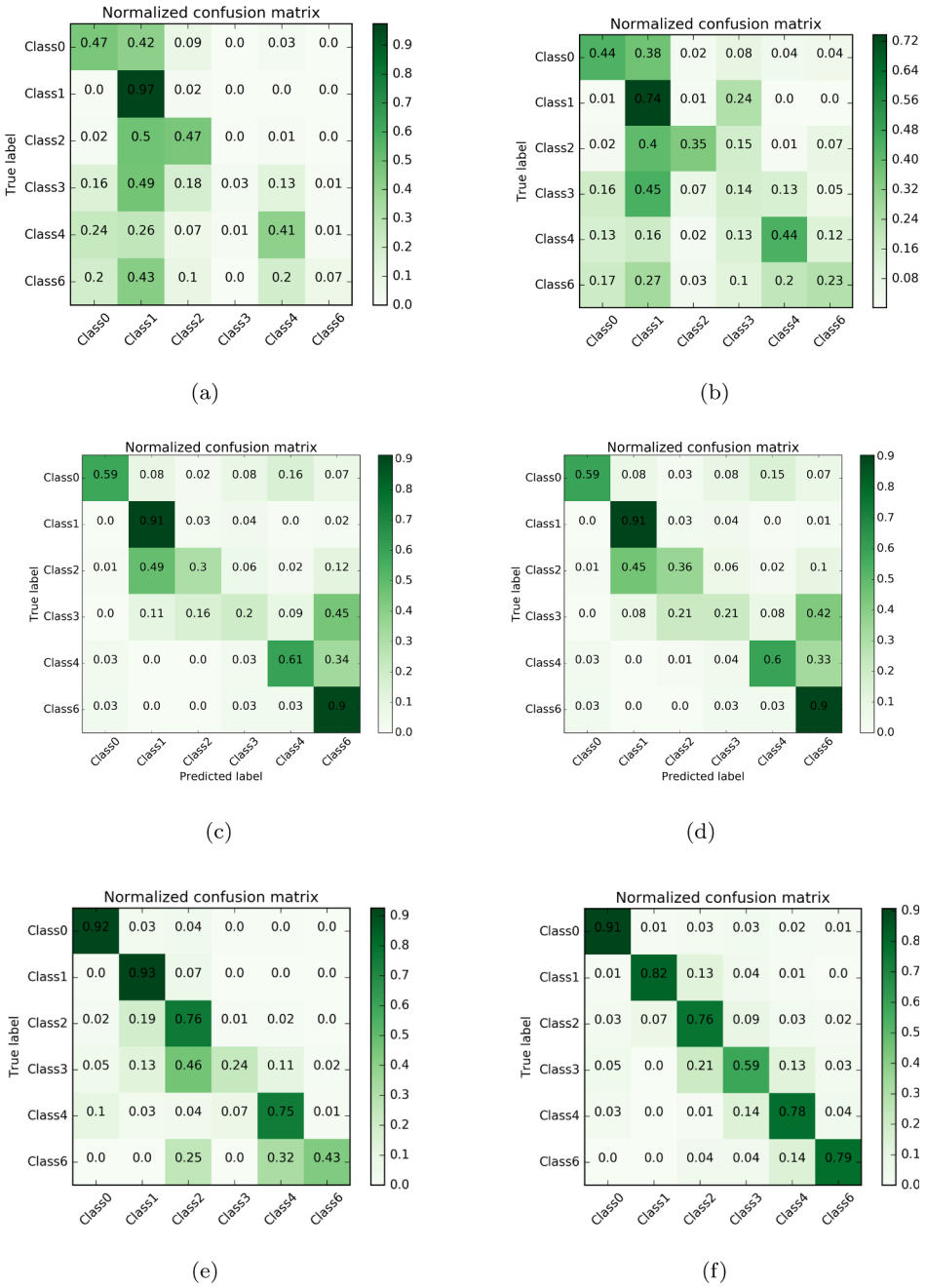


Fig. 5. Normalized confusion matrix for 4419 test RadTF reports – (a) rule-base method without sampling, (b) rule-base method with sampling, (c) out-of-the-box Word2Vec without sampling, (d) out-of-the-box Word2Vec with sampling, (e) proposed method without sampling, (f) proposed method with sampling. The numeric value in the diagonal cells represents the true positive rate.

trained on the proposed embedding is able to properly learn the malignant features of the radiological findings. Interestingly, the out-of-the-box Word2Vec model out-performed the rule-based system in some situations, which shows the power of unsupervised embedding to capture the semantics of radiology reports that not only saves tedious manual feature engineering, but also aids model generalizability.

To present the class-level comparative performance in a more interpretable manner, we visualize the corresponding confusion matrices in Fig. 5 that are computed in six distinct settings: (a) rule-based without sampling; (b) rule-based with sampling; (c) out-of-the-box Word2Vec model without sampling; (d) out-of-the-box Word2Vec model with sampling; (e) proposed model without sampling; (f) proposed model with sampling. The numeric values in the confusion matrices show the class-level classification agreement where 1 signifies that 100% of the data in BI-RADS class x is correctly classified and 0 signifies that no single report is classified. As seen from the figures (Fig. 5(a)), without sampling, the rule-based method was only able to derive good classification performance for BI-RADS class 1, with 97% of reports being correctly classified. This performance may be influenced by the strong skewed distribution of class 1 in the training dataset. Even after sampling (Fig. 5(b)), the accuracy of the rule-based method for inferring malignant and intermediate BI-RADS classes stayed as low as 33% and 14%, when in fact the proposed model's average accuracy for inferring malignant BI-RADS (BI-RADS classes 4 and 6) is 79%. The performance of out-of-the-box Word2Vec is better than the rule-based system. Mainly, it successfully classified BI-RADS class 0 and 1, as well as class 4 and 6, with agreement of 60%. Yet it failed to classify the intermediate BI-RADS classes 2 and 3. This is probably due to the

Table 5
Sample reports of BI-RADS 1 category and BI-RADS 6 category. (Finding sections only).

BI-RADS Class 1 – Limited intra-class language variation
The breast tissue is composed of scattered areas of fibroglandular density. No new focal dominant mass architectural distortion or suspicious microcalcifications are identified. There are no features to suggest malignancy.
The breast tissue is heterogeneously dense which may obscure detection of small masses. There is no new focal dominant mass architectural distortion or suspicious microcalcifications. There are no suspicious features to suggest malignancy.
BI-RADS Class 6 – Significant intra-class language variation
The breast tissue is almost entirely fatty. There is a post needle biopsy S-shaped metal marker in the lower and inner right breast which appears separated by 1.7 cm caudally from a small residual microcalcification with adjacent small hematoma.
The breast tissue is composed of scattered,areas of fibroglandular density. Three scar markers are seen within the right breast in the upper and upper outer quadrants associated with architectural distortion and compatible with postsurgical changes. An omega shaped marker with a small amount of high density tissue and a small group of microcalcifications are seen within the right breast at the 8:00 position 10.5 cm from the nipple at the site of previously biopsy-proven invasive ductal carcinoma. The soft tissue component appears smaller than on prior likely related to interval resolution of post-biopsy hematoma. The calcifications span a 15 mm region around the omega shaped marker. Other calcifications appear scattered and loosely grouped throughout the right breast.

semantic similarity in the language between the intermediate classes. For the proposed model, the sampling strategy helps in boosting the classification accuracy of the intermediate and malignant BI-RADS classes (see Fig. 5(d)). In most cases, the proposed model out-performed the rule-base method and the out-of-the-box Word2Vec model.

The failure of the rule-based method in categorizing the malignant BI-RADS classes is likely due to the fact that the intermediate and high-risk mammography reports (BI-RADS categories 3, 4, and 6) exhibit large variability in the text, which resulted in significant deterioration in the classifier performance using the rule-based method compared with the proposed context-aware embedding. On the other hand, the true semantics of the BI-RADS 0, 1, and 2 mammography reports (which indicate incomplete imaging or a low risk of findings) was adequately extracted by the rule-based method, since the low-risk reports are written using macros or very similar language with less term variation. The findings sections of four diagnostic mammogram reports are selected to illustrate the major language difference in high-risk reports (BI-RADS 6) compared to low-risk (BI-RADS 1) (see Table 5).

The classification performance of the BI-RADS class 3 is low for both methodologies: the best accuracy of the rule-based method is 14% while proposed method had 59% accuracy. This was due to the fact that there are many cases where the impression section contains critical information for BI-RADS categorization task, while the rest of the report, including the finding section, contains an inadequate description of the abnormality. Thus, the content of the reports without the impression section is insufficient for the BI-RADS classification task in these cases. This observation particularly holds for the BI-RADS class 3 (see Fig. 6). Interfering BI-RADS 3 is also challenging (sometimes with random assignment) for the human reader [29].

In Fig. 7, we present the 95% confidence interval for both the rule-based method and the proposed system (IWE), which is derived using an empirical bootstrap approach repeated 1000 times with a random number of samples. The bootstrap methodology can provide a decent estimate of how the model's accuracy might vary for a unseen sample set. The figure shows that though the spread of the confidence interval for both methods is small, and the mean accuracy of the proposed model is much higher than the rule-based system both with and without sampling.

4.2. External validation: oncoshare screening and diagnostic mammograms

On 1900 mammogram reports from Corpus 3, we applied both the rule-based and the proposed model, which have been trained only on Corpus 1 reports (radTF dataset). This validation strategy can provide insight about the generalizibilty of the models. We ran the validation with and without the sampling and present the results of the best performing models. The rule-based performed best without sampling the training set, whereas the proposed method performed best after sampling the training dataset. However, as seen from the classification results in Table 6 and the confusion matrix in Fig. 8, our proposed model successfully generalized on the Oncoshare reports with an average F1-score of 0.89, while the average F1-score for the rule-based system was 0.70. More importantly, despite being trained on a highly skewed dataset, the proposed system was able to accurately classify both high (BI-RADS 4,6) and low (BI-RADS 0,1,2) BI-RADS category, while the rule-based system mostly failed to classify the BI-RADS 3 and 6 reports.

In order to demonstrate the inter-observer variability for the BI-RADS inference task, we conducted an experiment with two independent radiologists (experienced with reading mammogragm reports) by randomly selecting 50 reports from Corpus 3. The radiologists were asked to read only the findings section of the mammogram reports and assign a BI-RADS score. We also annotated the same reports with the proposed model and present both results in Fig. 9, where the x-axis represents the individual report and the y-axis shows the assigned BI-RADS score. Between rater 1 and the original mammogram readers, there is an overall 22% disagreement, and there is 10% disagreement between rater 2 and the original mammogram readers. Interestingly, disagreement between the proposed model and the original mammogram readers is only 8%, which is lower than the radiologists reading the mammogram findings.

5. Discussion

The aim of our study was to propose an efficient and generalizable approach for inferring the BI-RADS final assessment categories by analyzing the clinical findings reported in free-text mammography reports. We proposed a hybrid semi-supervised method which combines semantic term embeddings with distributional semantics and showed

comparison: digital screening mammogram. technique:
lateral full-view and mag-spot, as well as mag-spot cc mammograms of the left breast with cad to aid in interpretation.
findings: three loosely grouped calcifications in the inner left breast were seen on mag-views.
impression: 1. left breast: bi-rads 3: probably benign. three loosely grouped calcifications were seen in the inner left breast on mag-views. diagnostic mammogram of the left breast is recommended in six months to establish stability of these lesions.

Fig. 6. Sample report with a BI-RADS 3 category. (Patient IDs, demographic characteristics, and dates have been redacted to preserve anonymity).

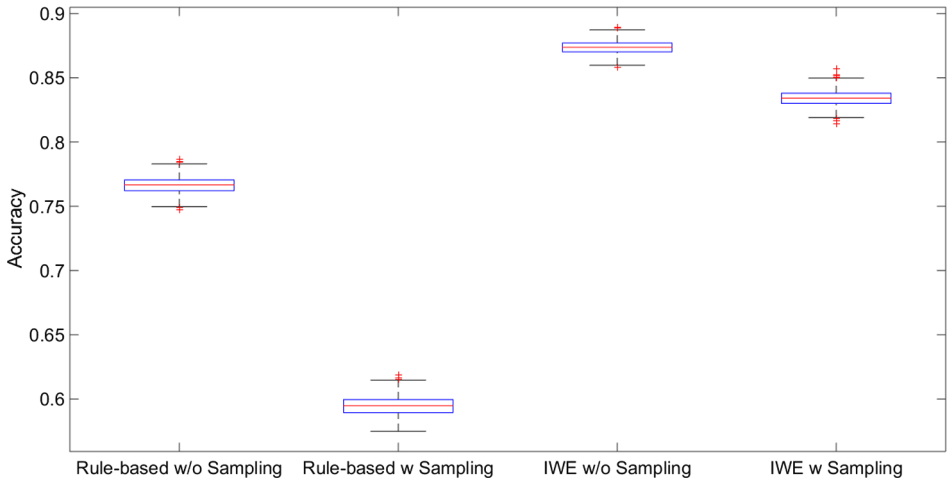


Fig. 7. Comparison of classification confidence between rule-based and proposed system (IWE).

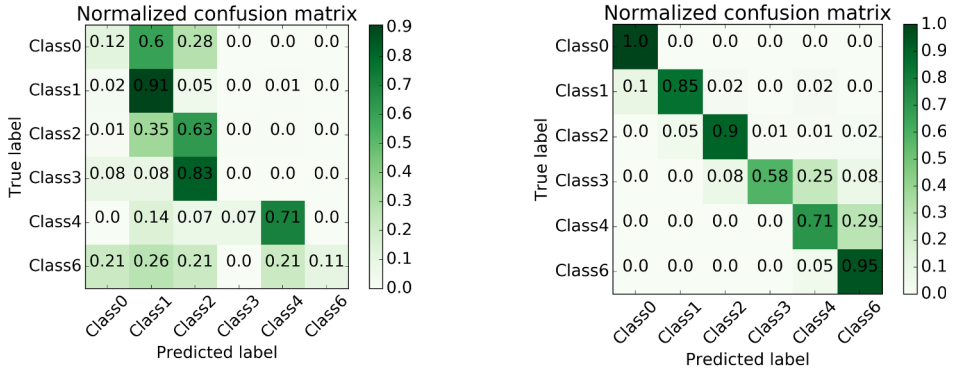
Table 6
Classification results for 1900 Oncoshare reports – proposed method against rule-based BN model.

	Corpus 3: OncoshareDB					
	Rule-based BN			Proposed method		
	Precision	*Recall*	*f1-score*	*Precision*	*Recall*	*f1-score*
Class 0	0.09	0.12	0.10	0.10	1.00	0.18
Class 1	0.32	0.91	0.48	0.87	0.85	0.86
Class 2	0.96	0.63	0.76	0.98	0.90	0.94
Class 3	0.00	0.00	0.00	0.70	0.58	0.64
Class 4	0.43	0.71	0.54	0.53	0.71	0.61
Class 6	0.33	0.11	0.16	0.67	0.95	0.78
Avg	0.84	0.66	0.70	0.91	0.87	0.89

that our methodology was able to analyze the free-text clinical findings and infer BI-RADS categorization with a performance comparable to the radiologist's reference standard. Semantic term embedding was employed to reduce the term-variations, as well as to create a vector representation of the whole report by abstracting significant radiological findings. To our knowledge, this is the first study that investigates the integration of distributional semantics techniques with the semantic term mapping for analyzing mammography reports. We also created a BI-RADS domain ontology automatically by developing a remote SPARQL API that extracts a BI-RADS specific sub-tree from the RadLex lexicon based on a set of terms provided by experts.

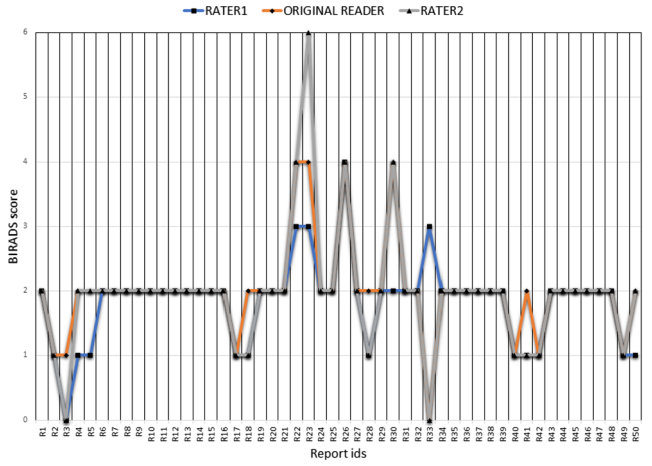
In terms of automatic classification of radiology reports, several recent studies have yielded promising results [11,6,12,8,30,10,31]. Nonetheless, the main limitation of the earlier works is that they require the domain experts to define concise information extraction rules or domain-specific terms for the free-text clinical narratives. Nguyen and Patrick [30] proposed active learning (AL) solutions for automatic feature selection, yet their system still needed to use some rule-based components. Analysis of free text narratives using rule based techniques is highly time-consuming and the rules are required to be reformulated for generalizability. Researchers have started to seek unsupervised or semi-supervised approaches with the help of recent development in NLP techniques. Neural word embeddings are one of the few currently successful applications of unsupervised learning. Their main benefit is that they do not require expensive text annotation, and the features can be derived from large unannotated corpora that are readily available in a domain such as radiology. Pre-trained embeddings can then be used in downstream tasks that use small amounts of labeled data.

In the current cohorts of reports that we used, BI-RADS final assessment categories are often explicitly mentioned in the impression section of the mammography reports. Therefore, we ran our pipeline considering only the findings sections. We also experimented with an external dataset (OncoshareDB) that combines mammogram reports from three different institutions and showed that our model, trained on a single institutional dataset, was able to performed well on these reports. These trials showed that our approach with domain knowledge formalization and context-based analysis is able to classify reports, even when no interpretation such as 'benign', 'BI-RADS 2', etc. occurs in the

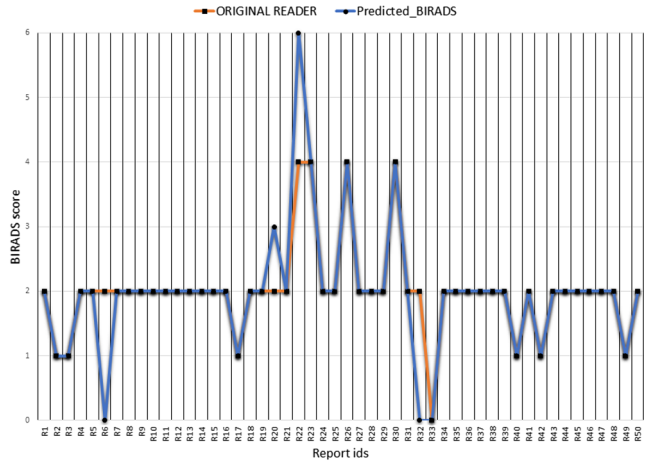


(a) (b)

Fig. 8. Normalized confusion matrix for 1900 Oncoshare reports – (a) rule-base method without sampling, (b) proposed method with sampling. The numeric value in the diagonal cells represents the true positive rate.



(a)



(b)

Fig. 9. Plot showing BIRADS scoring of 50 randomly pulled reports from Corpus 3 (a) assigned by the original mammographer and two individual radiologists; (a) assigned by the original mammographer and our model.

report text. The results also showed that our approach is useful in the identification of malignant cases. Even without any mention of BI-RADS classes in the findings section, report classification achieved promising results. This has the consequence of producing high specificity while maintaining reasonable sensitivity. On the other hand, like in example Fig. 6, most of the reports' findings sections were not complete in terms of imaging observations. We believe that the performance of the report categorization will be significantly improved if the findings section of the mammography reports comprehensively describe tumor characteristics. As suggested in the literature, more complete and more effective reporting of imaging observations [32] would promote the efficiency of NLP applications to any classification or extraction task.

The improvement in the performance of the standard classifier when compared with the rule-based system and the out-of-the-box Word2Vec model (see Section 4) suggests that the integration of semantic-dictionary based context analysis and unsupervised neural embedding can enhance the extraction of critical information from radiology reports by capturing the content of highly variable narration. We suspect that the low success rate of the rule-based system is also caused by the fact that the rule-based system was extracting information for each lesion in a mammography report and present the performance for assigning a lesion based BI-RADS category [13]. In contrast, in this study, our approach for was predicting BI-RADS categories on a document level. For instance, if there are more than one lesions in a report, all information was concentrated for all lesions in the document and the network was re-trained to classify document level final BI-RADS assessments. Therefore, concentration might be a reason for the performance drop of the rule based system comparing to its former results. In addition, we observed that the features that were extracted by the rule-based system were mostly sparse for both Corpus 1 and 3, especially for reports of the BI-RADS 3 class. This is due to the fact that the rule-based system was designed to find terms only if they are recorded in the findings section following the exact BI-RADS terminology. Linguistic variations and vague observations cannot be tackled by the rule-based system, which decreased the performance of the classification task. The out-of-the-box Word2Vec model performance is better than that of the rule-based system. This could be due to the fact that the Word2Vec model somewhat captured the linguistic variation by computing similarity between the words.

Efficient inference of BI-RADS classes may help to establish phenotyping in terms of malignancy for research purposes and can also expedite quality reporting for radiology practices. It is also concluded that if the term contexts are defined efficiently using embedding methods, such as we present in this work, then they can be used as features for other non-mentioned class terms in the text. Therefore, the technique might also be useful for automated coding and correction of reporting errors, and a second reading of reports to help increase the consistency of final assessments and decrease the ambiguities in the final assessments [33]. Our experiment shows that, on an external validation dataset, there is higher agreement between the original image reader and the model than two individual radiologists. There are a number of potential sources of variation for the individual radiologists, which include different personal ideas of what constitutes a truly negative exam vs. benign findings (BI-RADS 1 vs. 2), different levels of suspicion for equivocal findings (BI-RADS 2 vs. 3), uncertainty about whether a finding is truly biopsy-proven without access to the original electronic record (BI-RADS 6), and inferring from the report the nature of any technical limitation to imaging (BI-RADS 0).

Despite the apparent high accuracy of our system, our approach has several limitations. First, in addition to many reports with incomplete findings section, the imbalanced nature of our corpus could have biased the performance of the classification task. Second, instead of using only the BI-RADS related terms that are present in the RadLex ontology, it would be more effective to use all BI-RADS terms which are not yet formalized in RadLex. Third, due to an insufficient number of reports in BI-RADS 5 class for training, we ignored the class for this study, which limits the practical utility of this approach. However, given enough training data, the classifier model can easily be retrained. And finally, although we believe that our method is highly generalizable due to the success on the external validation set, our model is trained with single institutional dataset, and this might bias the performance of the classification task.

6. Conclusion

In conclusion, we have presented the first experimental demonstration of combining semantic context-driven analysis with a distributional semantics technique to classify mammography reports according to the BI-RADS classes. We believe that this technique can provide a valuable and efficient way for recognizing the BI-RADS assessment category phenotype in mammography reports without needing any pre-defined rules. Ultimately, our approach can automatically score reports describing breast cancer based on BI-RADS final assessment categories and may improve standardization in BI-RADS reporting. We believe the method may help to facilitate large scale text mining or data gathering tasks to improve decision making in breast cancer, and, most importantly, it requires minimal human effort for task-specific customization.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgement

This work was supported in part by grants from the National Cancer Institute, National Institutes of Health, 1U01CA190214, 1U01CA187947, and a grant from the Stanford-Philips Research Collaboration. The work was partially supported by the International Postdoctoral Research Fellowship Program Grant of the Scientific and Technological Research Council of Turkey (TUBITAK-2219). The authors acknowledge research support from the Breast Cancer Research Foundation and the BRCA Foundation.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103137.

References

[1] T. Samuels, Illustrated Breast Imaging Reporting and Data System Birads, American College of Radiology Publications, 1998.
[2] B. Boyer, S. Canale, J. Arfi-Rouche, Q. Monzani, W. Khaled, C. Balleyguier, Variability and errors when applying the birads mammography classification, Eur. J. Radiol. 82 (3) (2013) 388–397.
[3] W.A. Berg, C. Campassi, P. Langenberg, M.J. Sexton, Breast imaging reporting and data system: inter-and intraobserver variability in feature analysis and final assessment, Am. J. Roentgenol. 174 (6) (2000) 1769–1777.
[4] R. Ballard-Barbash, S.H. Taplin, B.C. Yankaskas, V.L. Ernster, R.D. Rosenberg, P.A. Carney, W.E. Barlow, B.M. Geller, K. Kerlikowske, B.K. Edwards, et al., Breast cancer surveillance consortium: a national mammography screening and outcomes database, AJR A. J. Roentgenol. 169 (4) (1997) 1001–1008.
[5] L. Liberman, J.H. Menell, Breast imaging reporting and data system (BI-RADS), Radiol. Clin. 40 (3) (2002) 409–430.
[6] 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 (e1) (2014) e81–e92.
[7] 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 (5) (2012) 913–916.
[8] C. Morioka, F. Meng, R. Taira, J. Sayre, P. Zimmerman, D. Ishimitsu, J. Huang, L. Shen, S. El-Saden, Automatic classification of ultrasound screening examinations of the abdominal aorta, J. Digital Imag. 29 (6) (2016) 742–748.
[9] I. Solti, C.R. Cooke, F. Xia, M.M. Wurfel, Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches, Bioinformatics and Biomedicine Workshop, 2009. BIBMW 2009. IEEE International Conference on, IEEE, 2009, pp. 314–319.
[10] G. Zuccon, A.S. Wagholikar, A.N. Nguyen, L. Butt, K. Chu, S. Martin, J. Greenslade, Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology, AMIA Summits Translational Sci. Proc. 2013 (2013) 300.
[11] S.M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, R.S. Jacobson, Automated annotation and classification of BI-RADS assessment from radiology reports, J. Biomed. Inform. 69 (2017) 177–187.
[12] D.A. Sippo, G.I. Warden, K.P. Andriole, R. Lacson, I. Ikuta, R.L. Birdwell, R. Khorasani, Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing, J. Digital Imag. 26 (5) (2013) 989–994.
[13] 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. Inform. 62 (2016) 224–231.
[14] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inform. Process. Syst. 2013, pp. 3111–3119.
[15] A. Gupta, I. Banerjee, D.L. Rubin, Automatic information extraction from unstructured mammography reports using distributed semantics, J. Biomed. Inform. (2018).
[16] H. Nassif, F. Cunha, I.C. Moreira, R. Cruz-Correia, E. Sousa, D. Page, E. Burnside, I. Dutra, Extracting BI-RADS features from portuguese clinical texts, Proceedings. IEEE International Conference on Bioinformatics and Biomedicine, NIH Public Access, 2012, p. 1.
[17] B.H. Do, A. Wu, S. Biswal, A. Kamaya, D.L. Rubin, Informatics in radiology: RADTF: a semantic search–enabled, natural language processor–generated radiology teaching file, Radiographics 30 (7) (2010) 2039–2048.
[18] S.C. Weber, T. Seto, C. Olson, P. Kenkare, A.W. Kurian, A.K. Das, Oncoshare: lessons learned from building an integrated multi-institutional database for comparative effectiveness research, AMIA Annual Symposium Proceedings, vol. 2012, American Medical Informatics Association, 2012, p. 970.
[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[20] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Syst. Man Cybernet. (3) (1972) 408–421.
[21] R. Řehuřek, P. Sojka, Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. < http://is.muni.cz/publication/884893/en > .
[22] S. Bird, NLTK: the natural language toolkit, Proceedings of the COLING/ACL on Interactive Presentation Sessions, Association for Computational Linguistics, 2006, pp. 69–72.
[23] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, Proc. GSCL (2009) 31–40.
[24] D.L. Rubin, D.A. Moreira, P. Kanjamala, M.A. Musen, BioPortal: a web portal to biomedical ontologies, AAAI Spring Symposium: Symbiotic Relationships Between Semantic Web and Knowledge Engineering, vol. 4, 2008, pp. 74–77.
[25] C.P. Langlotz, RadLex: a new method for indexing online educational materials (2006).
[26] S. Tamang, T. Hernandez-Boussard, E.G. Ross, M. Patel, G. Gaskin, N. Shah, Y. Bas, J.-F. Julien, D. Bas, J.G. Millichap, et al., Enhanced quality measurement event detection: an application to physician reporting, eGEMs (Generating Evidence & Methods to Improve Patient Outcomes) 5 (1) (2017) 5.
[27] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013, pp. 3111–3119.
[28] T. Kenter, A. Borisov, M. de Rijke, Siamese cbow: Optimizing word embeddings for sentence representations, arXiv preprint arXiv: < 1606.04640 > .
[29] A. Michaels, C. Chung, E. Frost, R. Birdwell, C. Giess, Interobserver variability in upgraded and non-upgraded BI-RADS 3 lesions, Clin. Radiol. 72 (8) (2017) 694–e1.
[30] D.H. Nguyen, J.D. Patrick, Supervised machine learning and active learning in classification of radiology reports, J. Am. Med. Inform. Assoc. 21 (5) (2014) 893–901.
[31] I. Banerjee, S. Madhavan, R.E. Goldman, D.L. Rubin, Intelligent word embeddings of free-text radiology reports, arXiv preprint arXiv: 1711.06968.
[32] E.S. of Radiology (ESR), Good practice for radiological reporting. guidelines from the european society of radiology (ESR), Insights Imag. 2 (2) (2011) 93–96.
[33] S. Bozkurt, D. Rubin, Automated detection of ambiguity in BI-RADS assessment categories in mammography reports, Stud. Health Technol. Inform. 197 (2014) 35–39.