Post-Structuring Radiology Reports of Breast Cancer Patients for Clinical Quality Assurance

Shreyasi Pathak, Jorit van Rossen, Onno Vijlbrief, Jeroen Geerdink, Christin Seifert, and Maurice van Keulen

Abstract— Hospitals often set protocols based on well defined standards to maintain the quality of patient reports. To ensure that the clinicians conform to the protocols, quality assurance of these reports is needed. Patient reports are currently written in free-text format, which complicates the task of quality assurance. In this paper, we present a machine learning based natural language processing system for automatic quality assurance of radiology reports on breast cancer. This is achieved in three steps: we i) identify the top-level structure (headings) of the report, ii) classify the report content into the top-level headings, and iii) convert the free-text detailed findings in the report to a semi-structured format (post-structuring). Top level structure and content of report were predicted with an F1 score of 0.97 and 0.94, respectively, using Support Vector Machine (SVM) classifiers. For automatic structuring, our proposed hierarchical Conditional Random Field (CRF) outperformed the baseline CRF with an F1 score of 0.78 versus 0.71. The determined structure of the report is represented in semi-structured XML format of the free-text report, which helps to easily visualize the conformance of the findings to the protocols. This format also allows easy extraction of specific information for other purposes such as search, evaluation, and research.

Index Terms—Quality assurance, automatic structuring, post-structuring, radiology reports, conditional random field

1 Introduction

Medical reports are essential for communicating the findings of imaging procedures with referring physicians, who further treat the patients by considering these reports. Since, medical reports are very important for diagnoses of diseases, there is a need to ensure that these reports are of a high quality. To maintain the quality of reports, hospitals often set well-defined protocols for reporting. For example, for breast cancer radiology reporting, hospitals generally use the "Breast Imaging-Reporting And Data System" (BI-RADS) [1], [20], which is a classification system proposed by American College of Radiology (ACR), to represent the malignancy risk of the breast cancer of a patient. It was implemented to standardize reporting and quality control for mammography. The BI-RADS lexicon provides specific terms to be used to describe findings. Along with that, it also describes the desired report structure: for example, a report should contain breast composition and a clear description of findings. The rate of compliance with these reporting standards can be used for quality assurance and also to further measure clinical performance [2].

Conformance to reporting standards can be seen as a part of assessing report clarity, organization, and accuracy [3], [4]. Quality assurance is currently mainly a manual process. Peer review is used to assess report quality, mainly geared towards accuracy of reports [5]. Yang et al. [6] used psychometric assessment to measure report quality and analyzed parameters like report preparation, organization, readability. Making quality assurance systems automatic would reduce the workload of radiologists and make the process more efficient. To the best of our knowledge, no system exists to automate this process.

Quality assurance is complicated due to the fact that reporting is done in free-text, narrative format. The inaccessibility of narrative structure for computers makes it hard to analyze if all the necessary information are present in the report. Structured reporting templates can be introduced to force the radiologists to stick to the reporting standards and improve the quality of reports [7], [8]. However, a study [9] shows that this type of system resulted in lower quality reports, as it restricts the style and format of writing. An alternative is to perform automatic structuring of free-text reports after they have been written, without additional technical burden on the radiologists. Thus, the radiologists can concentrate more on the task of interpreting images rather than structure of writing, which helps in maintaining accuracy of the report content.

In this work, we follow the post-structuring paradigm. We present an approach for automatic structuring of radiology reports for quality assurance using machine learning. We define the quality of a report by conformity to reporting standards set by ACR BIRADS. Concretely, we (i) identify the top-level structure from the reports (henceforth, referred to as heading identification), (ii) classify the report content into the top-level headings (referred to as content identification), and, (iii) automatically convert the free-text report

- *S. Pathak, C. Seifert, and M. van Keulen are with the Data Management and Biometric Group, University of Twente, Enschede 7522, NB, the Netherlands. E-mail: shreyasi12dgp13@gmail.com, {c.seifert, m.vankeulen}@utwente.nl.*
- *J. van Rossen, O. Vijlbrief, and J. Geerdink are with the Hospital Group Twente (ZGT), Hengelo 7555, DL, the Netherlands. E-mail: {j.vrossen, o.vijlbrief, J.Geerdink}@zgt.nl.*

Manuscript received 6 Oct. 2018; revised 30 Dec. 2018; accepted 4 Mar. 2019. Date of publication 3 May 2019; date of current version 8 Dec. 2020. (Corresponding author: Shreyasi Pathak.)
Digital Object Identifier no. 10.1109/TCBB.2019.2914678

TABLE 1
Example of Structuring of Free-Text Mammography Finding

Free-text Report	Structured Report
Mammografie t,o,v, 12/08/2016: Mamma compositiebeeld C, Geen wijziging in de verdeling van het mammaklierweefsel, Hierin beiderzijds geen haardvormige laesies, Geen distorsies, geen stellate laesies, geen massa's, bekende verkalking links, Geen clusters microkalk, geen maligniteitskenmerken,	⟨report⟩ ⟨O⟩Mammografie t,o,v, 12/08/2016:⟨/O⟩ ⟨breast_composition⟩Mamma compositiebeeld C,⟨/breast_composition⟩ ⟨O⟩Geen wijziging in de verdeling van het mammaklierweefsel,⟨/O⟩ ⟨negative_finding⟩ ⟨mass⟩Hierin ⟨location⟩beiderzijds⟨/location⟩ geen haardvormige laesies⟨/mass⟩ ⟨architectural_distortion⟩Geen distorsies,⟨/architectural_distortion⟩ ⟨mass⟩geen ⟨margin⟩stellate⟨/margin⟩ laesies, geen massa's, ⟨/mass⟩ ⟨/negative_finding⟩ ⟨positive_finding⟩ ⟨calcification⟩bekende verkalking ⟨location⟩links⟨/location⟩ ⟨/calcification⟩ ⟨/positive_finding⟩ ⟨negative_finding⟩ ⟨calcification⟩Geen ⟨distribution⟩clusters⟨/distribution⟩ ⟨morphology⟩microkalk,⟨/morphology⟩ ⟨/calcification⟩⟨/negative_finding⟩ ⟨O⟩geen maligniteitskenmerken⟨/O⟩ ⟨/report⟩

findings to a structured format for making the task of comparison to well-defined protocols easier (referred to as automatic structuring). For visualization and use in subsequent applications, we generate an output in a semi-structured XML format (Table 1). In this work, we focus on Dutch radiology reports on breast cancer; the automatic structuring was performed on findings from the mammography imaging modality in these reports. This article is an extended version of our previous work [10]. Among others, it adds more depth into error analysis of our task and additionally provides experiments with various feature combinations for the classifiers.

In the remainder of this paper, we first review structured reporting initiatives and natural language processing for radiology reports (Section 2). Next, we describe our data set in Section 3. Our approach to heading and content identification, and automatic structuring is detailed in Section 4, followed by the experimental setup (Section 5) and results (Section 6). Finally, we discuss the implication of our results in Section 7 and conclude our work in Section 8.

2 Related Work

In this section, we will discuss structuring initiatives for radiology reporting, and review work on natural language processing techniques in the domain of radiology.

2.1 Structured Reporting Initiatives

Accuracy, clarity, readability, and organization are some of the important factors for good quality of radiology reporting [3], [4]. Sistrom and Langlotz [7] identified i) language, ii) format as two key attributes for improving the quality of a radiology report. *Standardizing the language* of the report promotes common interpretation of the reports by the radiologists throughout the world. Breast Imaging-Reporting and Data System is a very successful attempt by ACR at standardizing the language for breast cancer reporting [1]. RadLex [11] is another attempt at standardizing disease terminology, observation and radiology procedure. *Structured reporting* further increases efficiency of information transfer and referring clinicians can extract the relevant information easily. Sistrom and Langlotz [7] clarified that structured reporting does not mean having a point-and-click interface for data capture, rather a simple report format that reflects the way radiologist and referring physician sees the report and should not impose any restriction on the radiologists. Radiological Society of North America (RSNA) highlighted that structured reporting would improve clinical quality and help in addressing *quality assurance* [4].

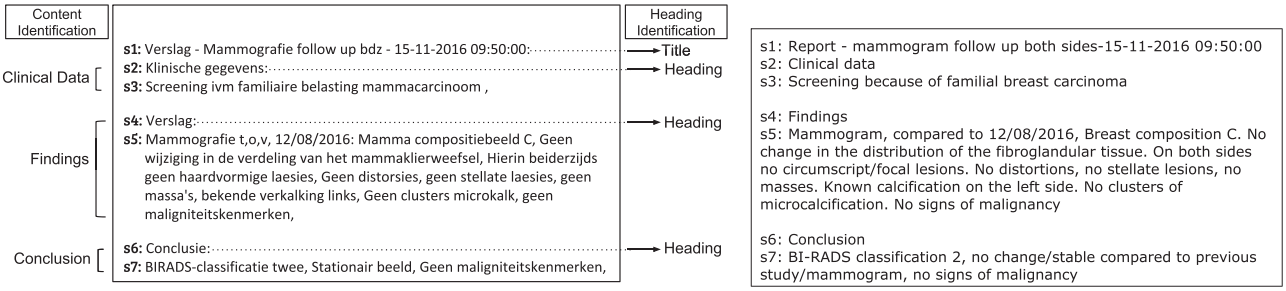
Though there has been a lot of discussion about the effect of structuring on the quality of radiology report, not much actual assessment was done until 2005. In 2005, Sistrom and Honeyman-Buck [12] tested information extraction from free-text and structured reports. They found that both, the free-text and structured report resulted in *similar accuracy and efficiency* in information extraction, but a post-experimental questionnaire expressed clinicians' opinion in favour of structured report format. Schwartz et al. [8] reported that referring clinicians and radiologists found *greater satisfaction with content and clarity* in structured reports, but the clinical usefulness did not vary significantly between the two formats. Another study by Johnson et al. [9], concluded that structured reporting resulted in a *decrease in report accuracy and completeness*. The subjects were asked to use commercially available structured reporting system (SRS), a point-and-click menu driven software, to create the structured reports and they found it to be *overly constraining* and *time-consuming*.

To summarize, previous work has shown that structured reporting and standard language are important for report quality, but should not impose restriction on the radiologists. Further, structured reporting can help in addressing quality assurance.

2.2 Natural Language Processing in Radiology

Electronic health records (EHRs), like radiology reports, increases the use of digital content and thus generates new challenges in the medical domain. It is not possible for humans to analyze this huge amount of data and extract relevant information manually, so automated strategies are needed. There are two types of techniques used in natural language processing for processing data: *i) rule-based* and *ii) machine learning* approaches.

In *rule-based approaches*, rules are manually created by experts to match a specific task. Various rule-based systems have been used for information extraction in breast cancer radiology reports. Nassif et al. [13] developed a rule-based system in 2009 to extract BI-RADS related features from a mammography study. The system was tested on 100



(a) Dutch radiology report (b) English translation of the Dutch report

Fig. 1. Example of a breast cancer radiology report.

radiology reports labeled by radiologists, resulting in a precision of 97.7 percent and a recall of 95.5 percent. Sippo et al. [14] developed a rule-based NLP system in 2013 to extract the BI-RADS final assessment category from radiology reports. They tested their system on >220 reports for each type of study – diagnostic and screening mammography, ultrasound etc. achieving a recall of 100 percent and a precision of 96.6 percent.

Machine learning (ML) approaches can learn the patterns from data automatically given the input text sequence and some labeled text samples. *Hidden Markov Model*, *Conditional Random Field (CRF)* [15] are some of the ML approaches used for sequence labeling. Hassanpour and Langlotz [16] compared dictionary-based (a type of rule-based) model, Conditional Markov Model and CRFs on the task of information extraction from chest radiology reports, finding that ML approaches (F_1: 85.5%) performed better than rule-based (F_1: 57.8%). Torii, Wagholikar and Liu [17] investigated the performance of CRF taggers for extracting clinical concepts and also tested the portability of the taggers on different datasets. Esuli, Marcheggiani and Sebastiani [18] developed a cascaded 2-stage Linear Chain CRF model (one CRF for identifying entities at clause level and another one at word level) for information extraction from breast cancer radiology reports. The cascaded system (F_1: 0.873) outperformed their baseline model of standard one level LC-CRF (F_1: 0.846) on 500 mammography reports.

Hybrid approaches combine rule-based and machine learning approaches. For example, Taira, Sodrland and Jakobovits [19] developed a method for automatic structuring of free-text thoracic radiology reports using some rule-based and some statistical and machine learning methods like maximum entropy classifier. We want to develop a fully automated system without any rule creation involved from experts, which is why we will not follow a hybrid approach.

In this work, we apply machine learning approaches to avoid manual rule construction and use CRFs, as they have shown high performance on sequence labeling task.

3 Corpus: Radiology Reports on Breast Cancer

According to BI-RADS [20], a breast cancer radiology report should contain an indication of examination (clinical data), a breast composition, a clear description of findings, and a conclusion with the BI-RADS assessment category. For our purpose of quality assurance of a report, we will consider these things and annotate the reports accordingly.

We used a dataset of 180 Dutch radiology reports on breast cancer from 2012 to 2017 (30 reports per year). Thus, the dataset contains variation in reports over the years. The reports were gathered from Hospital Group Twente (ZGT) in the Netherlands. The reports are produced by dictation from the radiologists, into an automatic speech recognition system. These automatically generated reports are further cross-checked with the dictation, by radiologists or secretary. The reports contain patient identity data like patient id, name, data of birth and address in separate columns, which were removed to anonymize the reports. A sample report is shown in Fig. 1a, with its English translation in Fig. 1b. The report has 3 sections, namely *Clinical Data*, *Findings* and *Conclusion*. *Clinical Data* contains clinical history of the patient including any existing disease or symptoms. *Findings* consists of noteworthy clinical findings (abnormal, normal) observed from imaging modalities like mammography, MRI and ultrasound. *Conclusion* provides a summary of the diagnosis and follow-up recommendations and should necessarily contain a BI-RADS category. In the report, these sections start with a heading describing the name of the section, for example, *Klinische gegevens* (Clinical Data), *Verslag* (Findings) and *Conclusie* (Conclusion). Reports from 2017 and 2016 (60 reports) additionally contain a *title*. The dataset consists of both male and female breast cancer reports; for automatic structuring, we focus on female reports.

For the first two sub-tasks of heading identification and content identification, 180 reports were manually annotated at the sentence-level by a trained expert. The reports were split into sentences, where for our data set, it was observed that a sentence means start of a new line, resulting in 1,591 sentences in total. In Fig. 1a, sentences are indicated by the labels s1 to s7. For the first sub-task of heading identification, sentences were labeled as *heading* (e.g., s2, s4, s6), *not heading* (e.g., s3, s5, s7) and *title* (e.g., s1). For the second sub-task of content identification, sentences were labeled as *title*, *clinical data* (e.g., s2, s3), *findings* (e.g., s4, s5) and *conclusion* (e.g., s6, s7). For the third sub-task of automatic structuring, we manually extracted the mammography imaging modality findings from the *findings* section of the report, which generated 108 mammography findings. These were manually annotated by two radiologists – a trainee (2 years of experience) and a consultant. Out of 108 reports, 18 reports were labeled collaboratively by both, 44 reports by the trainee and 46 by the consultant. After labeling, these 44 reports and 46 reports were analyzed to highlight any inter-annotator discrepancy, which were further resolved by the annotators.

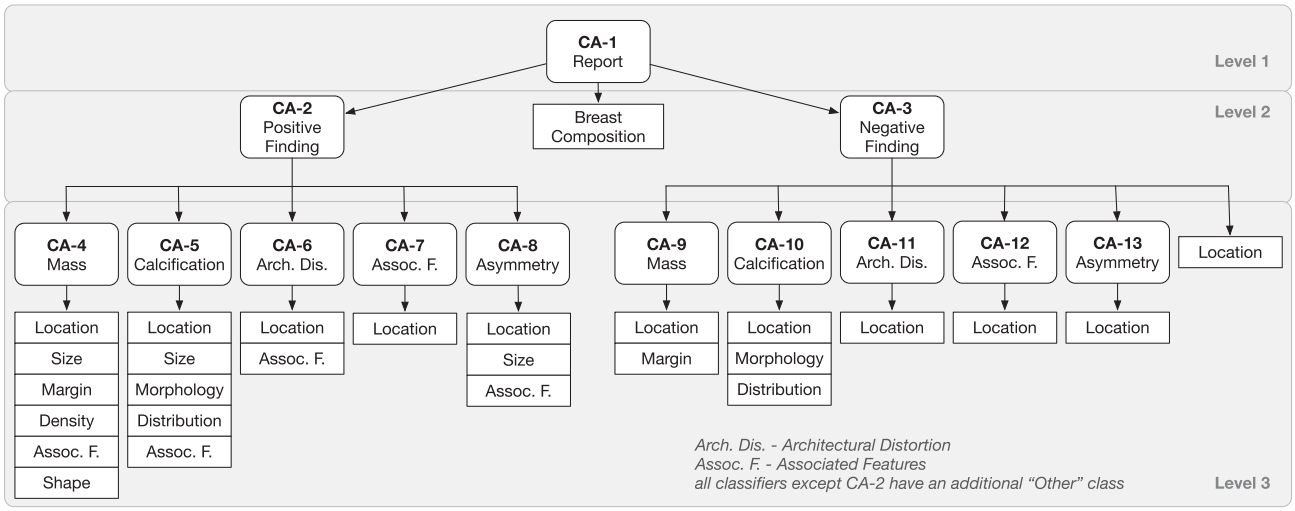


Fig. 2. 3-level annotation scheme for automatic structuring of mammography findings (Hierarchical Conditional Random Field Model A (Section 4.3.4)).

A 3-level annotation scheme at word-level was followed for automatic structuring as shown in Fig. 2. CA-n in the diagram will be explained in the approach (Section 4.3). At the first level, the reports were annotated as:

- *positive finding* (PF): something suspicious was detected about the lesion in the breast, which might indicate cancer.
- *negative finding* (NF): nothing bad was found or absence of specific abnormalities.
- *breast composition* (BC): density of the breast.
- *other* (O): text not belonging to the above.

After this first level of annotation, the PF were further annotated into second level classes – *mass* (MS), *calcification* (C), *architectural distortion* (AD), *associated features* (AF) and *asymmetry* (AS). At the third level, mass was further annotated as *location* (L), *size* (SI), *margin* (MA), *density* (DE), AF and *shape* (SH). Calcification was further annotated as *morphology* (MO), *distribution* (DI), SI, L and AF. Similar third level annotation was done with AD, AF and AS. The same scheme of second and third level annotation was followed for NF, though they have different combination of classes (as shown in Fig. 2). BC does not have any further levels of annotation. Thus, complete label (global) of a token is a concatenation of the labels at the 3 levels, resulting in 39 different labels. Our dataset only had data for 34 labels. This annotation scheme is based on the ACR BIRADS, with a few modifications done by our expert radiologists, e.g., a PF and a NF were added, a location class was added at the second level under NF to tag the location common for all the "no abnormalities", example, the phrase "left breast" in "no calcification, mass, architectural distortion was found in the left breast". Our model can also be applied to findings from other imaging modalities but it needs to be trained on manually labeled data for those modalities. Due to absence of labeled data from other modalities, we only performed automatic structuring of mammography findings.

4 Approach

In this section, we describe our approach for the three subgoals – heading identification, content identification, and automatic structuring of mammography findings.

4.1 Heading Identification

In this section, we describe the feature extraction and classifiers built for our task.

4.1.1 Feature Extraction

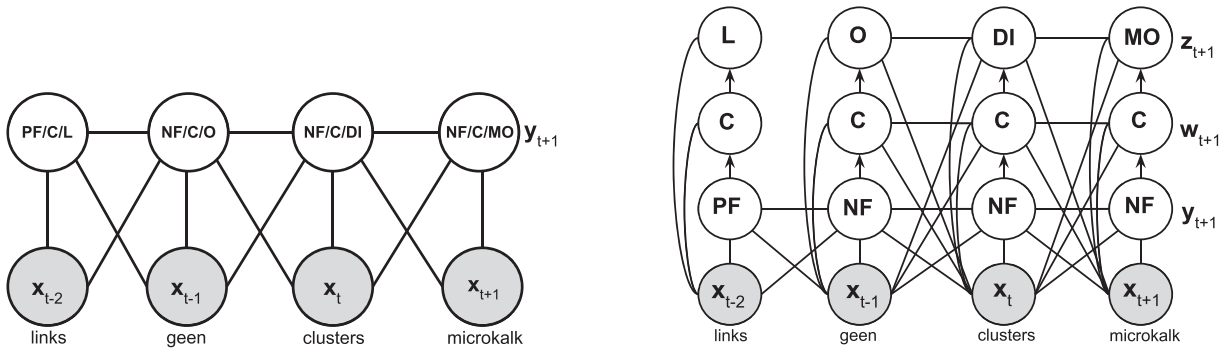
Reports were separated into sentences as explained in Section 3. The sentences were separated into word-level tokens using regular expression $\backslash b \backslash w \backslash w+ \backslash b$, which means tokens with at least 2 alphanumeric characters. Punctuations are always ignored and treated as token separator. For example, a sentence like *"Mammografie t,o,v, 12/08/2016: Mamma compositiebeeld C"* will generate {*mammografie, 12, 08, 2016, mamma, compositiebeeld*} as tokens. Only unigrams were taken as tokens and converted to lowercase. The maximum document frequency was set such that the terms occurring in more than 60 percent of the documents will be ignored. Increasing the maximum document frequency did not improve the performance, so most probably high frequency non-informative words were removed.

One of the features used was *word list feature*. A vocabulary was built using the unique words generated after preprocessing. Each sentence is represented by a term vector, where TF-IDF score is used for the tokens present in the sentence and a zero for absent tokens. The second feature is *length of sentence*, represented in two ways – i) number of tokens in the sentence and ii) logarithm to the base 10 of the value in (i) (this representation was used to get the length in the same value range as the other features). The third feature is the symbol at the end of sentence (EOS symbol). The headings end with a colon (:) usually and the rest of the sentences either end with a comma (,) or just a letter.

4.1.2 Classifiers

Heading identification is a multiclass classification problem, where the sentences are to be classified into one of the following classes: *heading*, *not heading* and *title*. We trained a Multinomial Naive Bayes (NB), a linear Support Vector Machine (SVM) and a Random Forest (RF) classifier.[1] For NB, Laplace smoothing was used. SVM was trained using stochastic

1. Classifiers were built using Python scikit-learn package.



(a) Baseline CRF model (b) Hierarchical CRF model

Fig. 3. Graphical representation for input feature vectors $\mathbf{x}_{t-2}$ to $\mathbf{x}_{t+1}$={links geen clusters microkalk}.

gradient descent and L2 loss. We used a maximum tree depth of 10 and bootstrap sampling for RF classifier.

4.2 Content Identification

Content identification is a multiclass classification problem, where the sentences are to be classified into *title*, *clinical data*, *findings* and *conclusion*. We followed the same approach as explained in Section 4.1, except, for feature extraction, we used only word list and length of sentence features. End of sentence symbol feature was not used, as sentences in two different sections usually end with similar symbol (','), thus, not contributing any unique feature to content identification problem.

4.3 Automatic Structuring

Our goal is to convert the free-text mammography findings into a semi-structured XML format. An example of this is shown in Table 1, where the first column shows a free-text mammography finding and the second column shows the semi-structured XML version. Let $\mathbf{X}$ be a mammography finding, consisting of a sequence of tokens, $\mathbf{x} = (x_1, x_2, .. x_t, .., x_n)$ and the task is to determine a corresponding sequence of labels $\mathbf{y} = (y_1, y_2, ..y_t, .., y_n)$ for $\mathbf{x}$. This task can be seen as *sequence labeling*, which is a task of predicting the most probable label for each of the tokens in the sequence. In this task, the context of the token, i.e., labels of immediately preceding or following tokens, is taken into account for label prediction. To achieve our goal, we used a Linear-Chain Conditional Random Field (LC-CRF)[2] [15], a supervised classification algorithm for sequence labeling. In our models, LC-CRF considers the label y_{t-1} of the immediately preceding token x_{t-1} for predicting the label y_t of the current token x_t.

4.3.1 Data Preprocessing

Each report from the dataset of 108 mammography findings was split at punctuations {,().?:-} (retaining them as tokens after splitting) and space, to generate tokens, $\mathbf{x}$, which were transformed according to the IOB tagging scheme [21]. Here, B means beginning of an entity, I means inside (also including end) of an entity and O means not an entity. For example, as shown in Table 1, "*Mamma compositiebeeld C,*" labeled as *breast_composition* was transformed to [(mamma, B-breast_composition), (compositiebeeld, I-breast_composition), (C, I-breast_composition), (',' , I-breast_composition)], where each entry stands for (token, label IOB scheme). Each digit was replaced by *#NUM* for the purpose of reducing the vocabulary size without removing any important information.

4.3.2 Feature Extraction

Each extracted token, x_t, is represented by a feature vector $\mathbf{x}_t$ for LC-CRF, including linguistic features of the current token, x_t, and also features of the previous token, x_{t-1}, and the next token, x_{t+1}. A feature vector $\mathbf{x}_t$ consists of the following 10 features for x_t and the same 10 features for x_{t-1} and x_{t+1} (a total of 30 features):

- The token x_t itself in lowercase, its suffixes (last 2 and 3 characters) and the word stem.
- Features indicating if x_t starts with a capital letter, is uppercase, is a Dutch stop word or is punctuation. The part-of-speech (POS) tag of x_t and its prefix (first 2 characters).

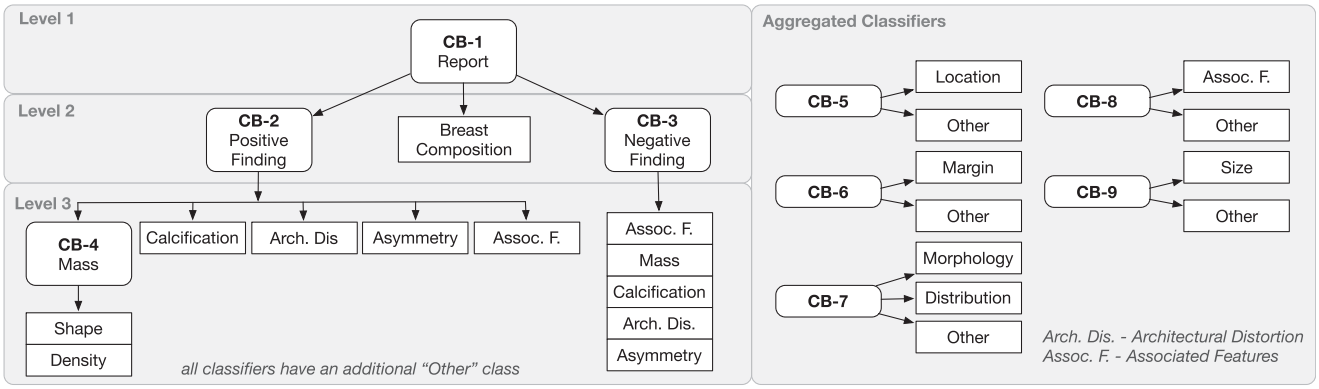
4.3.3 Baseline Model

As baseline, we used one LC-CRF classifier, as described at the starting of Section 4.3, to predict the complete label (concatenation of labels at the 3 levels) of a token and as input to the classifier, we used the feature vectors described in *Feature Extraction* (Section 4.3.2). For example, the LC-CRF classifier will predict the tokens *clusters* and *microkalk* as *NF/C/DI* and *NF/C/MO* respectively (see Table 1). The graphical representation of this model is shown in Fig. 3a. Here, $\mathbf{x}_{t-1}$, $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ are feature vectors of the tokens in a sequence and their corresponding labels are y_{t-1}, y_t, y_{t+1}, shown as NF/C/O, NF/C/DI, NF/C/MO. The lines indicate dependency on feature vectors $\mathbf{x}_{t-1}$, $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ and preceding label y_{t-1} for prediction of the label y_t. Thus, in this model, only one classifier is used to predict 34 labels.

4.3.4 Hierarchical CRF

We built a model using a three-level hierarchy of LC-CRF classifiers, called Model A, as shown in Fig. 2. The model has 13 LC-CRF classifiers and all the classifiers perform token-level prediction. One classifier (CA-1) is at level 1 for classifying the tokens into the first level classes. At level 2, there are 2 classifiers – one (CA-2) for further classifying the tokens predicted as *positive finding* by CA-1, another (CA-3)

2. We have used scikit-learn Python package, sklearn-crfsuite, implementation of LC-CRF.



Example: Positive Finding/Assymmetrie/Size is decided by classifier chain CB-1, CB-2, CB-9

Fig. 4. Hierarchical conditional random field model B.

for *negative finding* tokens. At level 3, there are 10 classifiers for further classification of tokens into third level classes. For example, the tokens classified as PF by CA-1 at level 1 and as MS by CA-2 at level 2, will be sent to CA-4 classifier to further get classified as either L, SI, MA, DE, SH or AF. The complete predicted label for each token is the concatenation of its predicted classes at the three levels. The graphical representation of this model is shown in Fig. 3b. For example, for given feature vectors $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ of the tokens *clusters* and *microkalk* respectively and for given classes at the same-level of the immediately preceding token, the first level class predictions for both the tokens are NF. The feature vector of these tokens are sent to NF classifier, CA-3, for second level prediction, where they get classified as C. Consequently, they are sent to the *calcification* classifier, CA-10, where they get classified as MO and DI respectively. Labels at each level are combined resulting in NF/C/DI and NF/C/MO labels for the two tokens. The undirected lines are dependency lines and directed lines are flow between the 3 levels (y, w, z). There is no dependency line between the first two columns at the second level (w) as *links* goes to PF and *geen* to NF classifier and two different classifiers are independent of each other's feature vectors and predicted class.

4.3.5 Hierarchical CRF with Combined Classes

As shown in Fig. 2, every classifier at level 3, predicts *location* as one of its classes. All the *location* classes describe similar tokens like *rechts, links, beide mamma*. Thus, we build one classifier for the similar classes instead of having different classifiers. This will provide us with more training data for a classifier. Fig. 4 shows the modified model with combined classes having 9 classifiers. Henceforth, this is referred to as Model B and all classifiers in this model are referred to as CB-n ($n = 1, \ldots, 9$). We can see instead of having 11 classifiers that predict *location* (CA-n, $n = 3, \ldots, 13$) in Model A, we have only one classifier CB-5 in Model B. Analogously, classifiers were aggregated for MA, MO, DI, AF and SI. All the classifiers use LC-CRF and perform token-level prediction. When classifying a token, classifiers might contradict each other. Consider for example NF/MS: CB-5 and CB-6 are the two classifiers predicting *location, margin* or *other* for the same token. If the predictions are *location* by CB-5 and *other* by CB-6, then *location* is selected (no contradiction). Similarly, if both classifiers predict *other*, then the resulting class is *other* (no contradiction). If the predicted class is *location* by CB-5 and *size* by CB-6 (contradiction), then the class with the highest a-posteriori probability is selected.

5 Experimental Setup

We used the F_1 score to evaluate the performance of a classifier on predicting different classes. The F_1 score of a class is the harmonic mean of precision and recall of that class and is defined as

$$F_1 = \frac{2TP}{2TP + FP + FN},$$

with TP, FP, FN being number of true positives, false positives and false negatives respectively. As our problem is a multiclass problem, the TP, FN, FP of a class are calculated according to one-versus-rest binary classification, where the class in consideration is positive and all other classes are negative.

We also measured F_1 score of the models on the entire test set using *micro-averaged* and *weighted macro-averaged* F_1 (F_1^μ and F_1^M). F_1^μ was computed by calculating the TP as sum over the TP of all the classes (same for FN, FP). F_1^M was calculated by computing the F_1 scores of each class separately and then averaging it. As, averaging gives equal weight to all the classes, the fact that our classes have unequal number of instances, is not taken into account. Thus, we used weighted averaging for F_1^M. F_1^M and F_1^μ gave similar results, so we only report F_1^M scores in the rest of the paper.

We evaluated how well the classifiers predict tokens as well as phrases. For phrases, we consider complete and partial matches. At the token-level (TL), we consider all the token labels in the dataset to calculate the TP, FP, FN scores of a class. At the partial phrase-level (PP) and the complete phrase-level (CP), we measure how well the classifier is performing in identifying multi-token phrases. A complete match requires all the tokens of the phrase to be correctly labeled. We consider a match with Dice's coefficient greater than 0.65 as a partial match. For similarity calculation, we take the phrase from the ground truth and match with the corresponding predicted labels. Phrase-level scores are important from the radiologists' point of view, as they care about how well their phrases match. Table 2a shows 6 tokens, with their token-level labels (B-PF, I-PF etc). A PF phrase starts at the B-PF and ends at the last I-PF. For the NF phrase, the Dice's coefficient is calculated as $2 * 2/(3 + 3) = 0.66 > 0.65$,

TABLE 2
Token Level and Phrase Level Measures

(a) Tokens and phrases

	bekende	verkalking	links	geen	clusters	microkalk
true	B-PF	I-PF	I-PF	B-NF	I-NF	I-NF
predicted	B-PF	I-PF	I-PF	O	B-NF	I-NF
true	PF phrase			NF phrase		
predicted	PF complete phrase match			NF partial phrase match		

(b) Token and phrase level scores

Classes	TL F_1	PP-Acc	CP-Acc	PP-Sim	#Tokens	#Phrases
BC	0.94	0.93	0.93	1.00	622	99
NF	0.95	0.97	0.91	0.99	1101	118
PF	0.87	0.87	0.87	1.00	1090	87

resulting in a partial match. For each class, we calculate the number of partial matches called partial phrase accuracy (PP-Acc); how well the partial phrases match by averaging the Dice's coefficient for each match (PP-Sim); the number of complete matches (CP-Acc); and the F_1 scores for token-level matches (TL F_1).

For heading and content identification, we evaluated NB, SVM and RF models, using 5-fold cross validation on 180 reports. We measured the performance of these classifiers for different combinations of features and analyzed for which feature combination the classifiers gave the best performance. Features used for heading identification are TF-IDF word list, EOS symbol and log length of the sentence; and for content identification – word list features represented in form of term frequency and TF-IDF, and length of the sentence feature represented in terms of number of tokens and log to the base 10 of number of tokens.

For automatic structuring, we built three different LC-CRF models: the baseline model, Model A and Model B. We evaluated our models using 4-fold cross validation on 108 mammography findings. For automatic structuring, we evaluated the models on different combinations of classes (Table 4c). 'All' means evaluation on all the 34 classes. 'w/o O' means all the classes except the *other (O)* class at the first level (33 classes). 'w/o <10&O' means classes excluding O class and classes with instances <10. Normalized confusion matrix, where all the values in each row of a confusion matrix are divided by the sum of all the values in that row (class support size) was calculated for automatic structuring task. All codes associated with this paper are available as open source.[3]

6 Results

In this section, we describe the results of heading and content identification and automatic structuring.

6.1 Heading Identification

Table 3a shows the performance of heading identification classifiers for different feature combinations. NB performed better with word list feature than with all the three features, whereas, SVM's performance did not change on adding EOS symbol and length feature on top of word list feature. Overall, it can be seen that word list itself is a very informative feature. EOS symbol feature was better informative than length of the sentence, as, for all the classifiers, the F_1^M for TF-IDF + EOS symbol is either same or better than TF-IDF + length. The scores in Table 4a are for the best feature combination i.e., the word list feature. It shows that the classes *headings* and *not headings* were predicted with an F_1 score of 0.96 and 0.98 respectively both by SVM and NB. For these classes, SVM and NB performed better than RF but for *title*, RF performed better. Fig. 5a shows the heat map representation of confusion matrix for heading identification using SVM and word list features. It can be seen that only 26 out of 540 heading instances were confused with not heading class.

6.2 Content Identification

Table 3b shows the performance of the content identification classifiers for different feature combinations. SVM shows the best performance ($F_1 = 0.94$) for TF-IDF word list feature and the scores in Table 4b are based on only this feature. In general, log length performs better than token length of sentence. The token length is a feature with high variance (short and long sentences), the log length varies much less. The token length of the sentence effected SVM much worse than NB as NB does an implicit normalization of features. Table 4b shows that SVM performed better for predicting the classes *conclusion*, *clinical data*, *title* and *findings* with an F_1 score of 0.92, 0.94, 0.99 and 0.94 respectively. Fig. 5b shows the heat map of confusion matrix for content identification using SVM classifier and word list feature. Both the conclusion and clinical data classes were wrongly predicted as findings in 44 and 25 out of their total instances respectively. This can be explained because conclusion, clinical data and findings, although being different, have similar words in their description, leading to the misclassification.

6.3 Automatic Structuring

Table 4c compares the performance of our baseline model to the hierarchical Models A and B. Both, Model A and B

TABLE 3
Performance of the Classifiers in Terms of F_1^M Scores for Different Feature Combinations

(a) Heading identification

Features	NB	SVM	RF
TF-IDF	0.97	0.97	0.92
TF-IDF + Length ($\log_{10}$)	0.93	0.97	0.94
TF-IDF + EOS Symbol	0.95	0.97	0.95
All Features	0.91	0.97	0.94

(b) Content identification

Features	NB	SVM	RF
Term frequency	0.91	0.92	0.79
TF-IDF	0.87	0.94	0.80
Term frequency + Length	0.87	0.40	0.81
TF-IDF + Length	0.70	0.29	0.82
Term frequency + Length ($\log_{10}$)	0.91	0.92	0.81
TF-IDF + Length ($\log_{10}$)	0.80	0.92	0.82

3. Source Code, https://doi.org/10.5281/zenodo.2717897

TABLE 4
Heading and Content Identification and Automatic Structuring Performance in Terms of F_1 Scores

(a) Heading identification

Classes	NB	SVM	RF	#Instances (Sentences)
Heading	**0.96**	**0.96**	0.88	540
Not Heading	**0.98**	**0.98**	0.94	991
Title	0.97	0.98	**0.99**	60
Avg (F_1^M)	0.97	0.97	0.92	1591

(b) Content identification

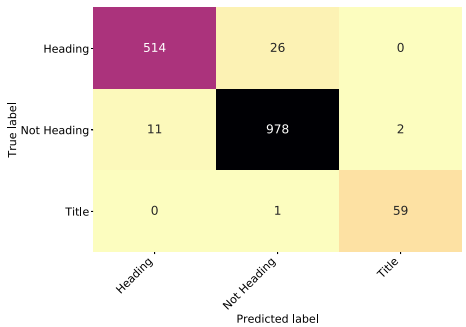
Classes	NB	SVM	RF	#Instances (Sentences)
Conclusion	0.89	**0.92**	0.90	413
Clinical Data	0.86	**0.94**	0.70	405
Title	0.89	**0.99**	0.91	60
Findings	0.88	**0.94**	0.82	678
Avg (F_1^M)	0.88	0.94	0.81	1556

(c) Automatic structuring

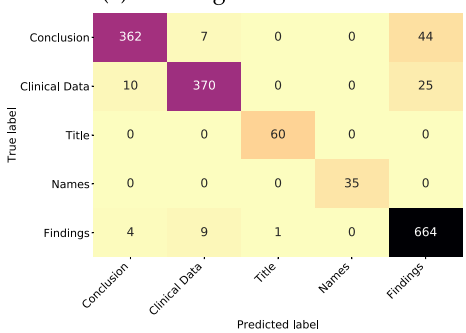
Measures	Baseline	Model A	Model B	#Instances (Tokens)
F_1^M(all)	0.71	0.78	0.78	4230
F_1^M(w/o O)	0.67	0.73	0.74	2813
F_1^M(w/o<10&O)	0.70	0.76	0.76	2649

(F_1^M = 0.78) outperformed the baseline model (F_1^M = 0.71). No difference in performance was observed between Model A and B. Without the not important *other (O)* class, the Model B has F_1^M = 0.74. On further removing classes with instances < 10, the F_1^M score improves from 0.74 to 0.76 for Model B. This means that the classes having instances < 10 were not predicted well enough. If we would have at least 10 instances for each class, then the F_1^M score could be expected to be around 0.76.

Table 2b shows the performance of the classifier (CA-1 and CB-1) at the first level in predicting *breast composition*, *negative finding*, *positive finding*. BC (TL F_1 = 0.94) and NF (TL F_1 = 0.95) were identified better than PF (TL F_1 = 0.87). This is because PF contains varied vocabulary for describing an abnormality, while NF contains specific terms like no presence of mass, calcification. BC is also described using specific terms like "mamma compositiebeeld". Token-level measure is always higher than complete phrase-level measure. PP-Acc is at least as good as CP-Acc. All the partial phrase matches in BC and PF are complete matches except for NF. But even for NF, the partial phrases have similarity of 0.99 (PP-Sim) with the ground truth.



(a) Heading identification



(b) Content identification

Fig. 5. Confusion matrix heat map for SVM classifier.

Table 5 shows the performance of the classes at the second level for the 3 models. Positive finding classifiers CA-2 and CB-2 at level 2 for Model A and B are similar and therefore, their F_1 scores are also same. But the negative finding classifier CA-3 and CB-3 are not similar for Model A and B, leading to different scores. The baseline model failed to predict the PF/AF and PF/AS classes but the hierarchical models successfully predicted the PF/AS class with 0.57 F_1 score and very weakly predicted PF/AF with an F_1 score of 0.11. PF/MS was predicted best among all the PF sub classes. There is a decrease in the overall PF sub classes prediction at the second level in comparison to the PF prediction at the first level for Model A and B. This shows even though PF class at the first level was predicted with good enough F_1 score of 0.87, the PF classifiers at the second level did more errors in predicting the second level PF classes. For the baseline model, as the global classes get predicted as a whole, it can be interpreted that F_1 score of 0.49 for PF classes at the first level was because of the PF/MS and PF/C sub classes. Among all the NF sub classes at level 2, NF/AF class was predicted the best (F_1 = 0.96) by the hierarchical models. From the dataset, it was found that NF/AF had a very similar sentence in all the reports, e.g. "Huid-subcutis geen bijzonderheden", leading to the high F_1 score. NF/L was at least slightly predicted by Model B, as Model B has an aggregated location classifier CB-5.

Table 6 shows the TL F_1 performance obtained for all the global classes. #Reports means the number of reports consisting of a class. #Phrases shows the number of phrases of each class. #Tokens contains the number of tokens belonging to a class and a phrase consists of multiple tokens – Each B-X, I-X are tokens for class X. Class 'O' was not labeled as B-X, I-X as

TABLE 5
Prediction of Second Level Classes in Terms of F_1 Score for the Three Models of Automatic Structuring

Classes	Baseline	Model A	Model B	Instances
PF/MS	0.53	0.66	0.66	483
PF/C	0.46	0.58	0.58	311
PF/AD	0.00	0.00	0.00	16
PF/AF	0.00	0.11	0.11	67
PF/AS	0.00	0.57	0.57	30
NF/MS	0.92	0.92	0.89	262
NF/C	0.88	0.85	0.88	260
NF/AD	0.89	0.90	0.88	77
NF/AF	0.96	0.96	0.96	403
NF/AS	-	-	-	-
NF/L	0.00	0.00	0.20	10
NF/O	0.89	0.82	0.79	88
Avg (F_1^M)	0.70	0.75	0.75	2007

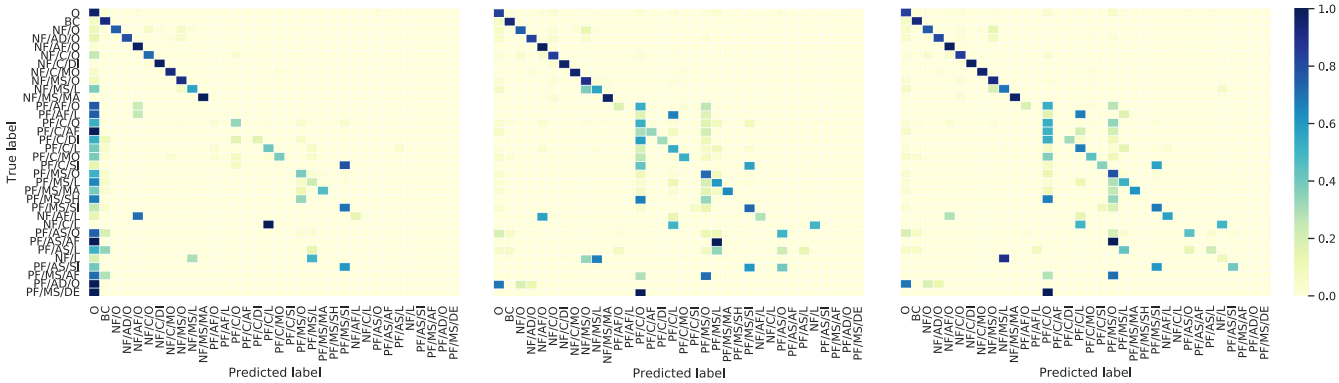
TABLE 6
Global Classes in the Dataset and Their F_1 Scores

Classes	#Tokens	#Phrases	#Reports	Baseline	Model A	Model B
O	1417	-	108	0.78	0.86	0.86
BC	622	99	97	0.89	0.94	0.94
PF/MS/L	139	33	27	0.29	0.40	0.47
PF/MS/SI	86	23	22	0.67	0.66	0.69
PF/MS/MA	59	22	20	0.53	0.72	0.70
PF/MS/DE	2	1	1	0.00	0.00	0.00
PF/MS/AF	7	2	2	0.00	0.00	0.00
PF/MS/SH	3	3	3	0.00	0.00	0.00
PF/MS/O	187	70	27	0.48	0.52	0.47
PF/C/L	68	38	35	0.49	0.44	0.59
PF/C/SI	14	5	5	0.00	0.00	0.22
PF/C/MO	39	37	32	0.52	0.56	0.51
PF/C/DI	19	13	11	0.25	0.58	0.53
PF/C/AF	33	6	6	0.00	0.17	0.00
PF/C/O	138	68	38	0.45	0.37	0.37
PF/AD/L	0	0	0	-	-	-
PF/AD/AF	0	0	0	-	-	-
PF/AD/O	16	1	1	0.00	0.00	0.00
PF/AF/L	6	6	5	0.00	0.00	0.00
PF/AF/O	61	11	7	0.00	0.12	0.13
PF/AS/L	35	14	11	0.00	0.14	0.17
PF/AS/SI	5	2	2	0.00	0.00	0.36
PF/AS/AF	1	1	1	0.00	0.00	0.00
PF/AS/O	172	13	11	0.00	0.58	0.56
NF/MS/L	17	14	13	0.60	0.50	0.50
NF/MS/MA	35	35	35	1.00	0.96	0.97
NF/MS/O	210	113	61	0.93	0.88	0.89
NF/C/L	2	1	2	0.00	0.00	0.00
NF/C/MO	56	56	51	0.95	0.91	0.97
NF/C/DI	54	53	50	0.98	0.98	0.99
NF/C/O	148	100	62	0.81	0.76	0.81
NF/AD/L	0	0	0	-	-	-
NF/AD/O	77	46	43	0.89	0.88	0.88
NF/AF/L	6	7	5	0.13	0.30	0.39
NF/AF/O	397	71	63	0.96	0.96	0.96
NF/AS/L	0	0	0	-	-	-
NF/AS/O	0	0	0	-	-	-
NF/L	10	6	6	0.00	0.00	0.20
NF/O	88	46	31	0.89	0.82	0.79
Total/Avg (F_1^M)	4229	1016	-	0.71	0.78	0.78

phrase of 'O' is not important, that is why there is no entry for phrases for class 'O'. PF/AD/L, PF/AD/AF, NF/AD/L, NF/AS/L and NF/AS/O does not occur in our dataset and that is why the values corresponding to them are 0. Overall, it can be seen that NF sub-classes were predicted better than PF sub-classes, as most of the NF sub-classes are described using specific tokens. Generally, Model A and B predicted PF sub-classes better than baseline. BC, NF/AF/O, NF/C/DI, NF/MS/MA and NF/C/MO were predicted very well in all the models. Some classes were predicted better in baseline – NF/MS/O, NF/MS/MA and PF/C/O. This indicates that for these classes, the neighbouring global classes of the baseline model may be informative during prediction. Also, multi-level prediction increased the number of false positives for a class, specially for classes with greater number of instances. The effect of aggregated classifiers in model B can be seen in NF/C/DI, NF/C/MO, PF/C/L, PF/MS/L and PF/C/SI. As the aggregated classifiers were trained on all L, DI, MO and SI in the dataset, it resulted in better prediction of third level classes like L, SI, even with few instances (14 tokens of PF/C/SI). But aggregating classifiers also resulted in loss of information about the context, which is reflected through slightly lower performance in Model B for classes PF/MS/MA, PF/C/AF and PF/AS/O. Aggregating AF classifier (CB-8) did not help in predicting any third level AF classes in PF due to not much similarity in their descriptions.

Fig. 6 shows the normalized confusion matrix heat map of global classes for baseline model, Model A and Model B. In baseline model, most classes were misclassified as other class and only BC and most NF classes were classified correctly. NF/C/L was wrongly predicted as PF/C/L, as location (L) of both NF and PF can be described in a similar manner. Similarly, PF/C/SI, PF/AS/SI were wrongly predicted as PF/MS/SI, as size (SI) of MS, C and AS are always written in a similar way in reports. For Model A and B (Fig. 6), there were not many misclassification with other class as for these models, tokens can only be misclassified into other class at the first level. In Model A and B, PF/MS/L were misclassified as PF/C/L, whereas in baseline, it was misclassified as other. Some other noteworthy observations between Model A and B are Model B had more true positives than Model A. Model A did not have any true positive of NF/L, PF/C/SI and PF/AS/SI whereas Model B had some. Model A misclassified PF/AS/SI as PF/AS/O, which shows misclassification at the third level. This observation proves that for Model B, aggregated classifiers like size helped in better prediction of third level classes.

Table 7 gives an indication of error propagation through the classifiers at the 3 levels for Model A and B. ΔF_1^M at a level indicate the difference in F_1^M of that level of classifiers on predicted classes when given true classes from previous level and when given predicted classes from previous level.



(a) Baseline model (b) Model A (c) Model B

Fig. 6. Normalized confusion matrices for automatic structuring.

TABLE 7
Error Propagation Through Classifiers at the Three Levels

Measures	Level2_A	Level2_B	Level3_A	Level3_B
ΔF_1^M	0.05	0.04	0.17	0.16
#Instances	2191	2191	2093	2093

This can be interpreted as error made by the classifiers at the previous level. Error made by level 1 (ΔF_1^M at level 2) is not much significant as compared to error by level 2 (ΔF_1^M at level 3) as the latter is a combination of errors from both level 1 and level 2 classifiers, while the former only considers error from level 1.

Fig. 7 shows comparison between the ground truth and predicted labels of a sample report (Table 1) for the task of automatic structuring using Model B. It can be seen that most of tokens were correctly predicted. Only one positive finding between the two negative findings got misclassified and combined with the negative finding.

7 Discussion

The first task of heading identification achieved a high F_1 score of 0.97 on TF-IDF word list features using SVM classifier. Adding features such as log length and end of sentence symbol did not change the F_1 score of SVM classifier. The second task of content identification achieved a high F_1 score of 0.94 on TF-IDF word list using SVM classifier. Adding length (in terms of number of tokens) as a feature hugely decreased the F_1 score ($F_1 = 0.29$) and adding log length just decreased it slightly ($F_1 = 0.92$).

For the third task of automatic structuring, the first level classes, breast composition and negative finding got predicted better than positive finding. We found out that breast composition and negative finding classes was described in a very specific way in the reports unlike positive finding, which was described using varied vocabulary. This made the prediction for positive finding harder than the other two.

On the second level, the positive finding classes mass and calcification were better predicted than asymmetry, associated features and architectural distortion. This is because far lesser training data was available for the latter classes. Further, from discussion with radiologists, we found out, that asymmetry findings are also hard to understand for humans. So, low scores on asymmetry can be expected. As negative finding class describes absence of abnormality using specific words e.g., no presence of mass or calcification, so, all the second level sub classes in negative finding – mass, calcification, architectural distortion, asymmetry and associated features, were predicted very well.

All the third level sub classes for negative finding were predicted very well compared to positive finding sub classes, due to better prediction of negative finding classes at first and second level. Morphology, distribution and margin are some of the third level sub classes with very high score. This is because morphology can be described using very specific words like "micro calcification" and "macro calcification", distribution can be described using "cluster" and margin can be described using words like "stellate" or "star-shaped". Among all the third level sub classes in positive finding, size and margin had the best results with F_1 score of 0.69 and 0.70 respectively as these were the classes most easily recognizable due to their specific format or words. Density and shape could not be recognized due to very little training data (around 2 or 3 tokens). Both second level and third level sub classes of associated features in positive finding were also very poorly recognized due to very less number of training data available.

Hierarchical models, Model A and Model B, did not vary significantly in overall performance. But, some classes were predicted better in Model B due to the use of aggregated classifiers. These were those classes, with similar description in all the groups and with less training data in each of these groups. So, the aggregated classifiers for these classes resulted in a lot of training data from the groups with that class, leading to better performance in Model B, e.g., classes like location and size. On the other hand, for some classes, better performance in Model A was observed than Model B. This is because information about the context of a token is available to classifiers of Model A. In Model A classifiers,

```
<report>
   <O>Mammografie t,o,v, 12/08/2016:</O>
   <breast_composition>Mamma compositiebeeld C,</breast_composition>
   <O>Geen wijziging in de verdeling van het mammaklierweefsel, </O>
   <negative_finding>
      <mass>Hierin
         <location>beiderzijds</location> geen haardvormige laesies,
      </mass>
      <architectural_distortion>Geen distorsies,
      </architectural_distortion>
      <mass> geen
         <margin> stellate</margin>laesies, geen massa's,
      </mass>
   </negative_finding>
   <positive_finding>
      <calcification> bekende verkalking,
         <location>links</location>
      </calcification>
   </positive_finding>
   <negative_finding>
      <calcification>Geen
          <distribution>clusters</distribution>
          <morphology>microkalk,</morphology>
      </calcification>
   </negative_finding>
   <O>geen maligniteitskenmerken,</O>
</report>
```

```
<report>
    <O>Mammografie t,o,v , 12/08/2016 :</O>
    <breast_composition>Mamma compositiebeeld C ,</breast_composition>
    <O>Geen wijziging in de verdeling van het mammaklierweefsel ,</O>
    <negative_finding>
       <mass> Hierin
          <location>beiderzijds</location>geen haardvormige laesies ,
       </mass>
       <architectural_distortion>Geen distorsies ,
       </architectural_distortion>
       <mass>geen
          <margin>stellate</margin> laesies , geen massa's ,
          bekende verkalking ,
          <location>links</location>
       </mass>
       <calcification>Geen
          <distribution>clusters</distribution>
          <morphology>microkalk ,</morphology>
       </calcification>
    </negative_finding>
    <O>geen maligniteitskenmerken ,</O>
</report>
```

Fig. 7. Automatic structuring: Comparison of the ground truth and the predicted labels by Model B for a sample report.

each token is surrounded by various classes in that group, e.g., associated features class is surrounded by distribution, morphology, location etc, in the group positive finding/calcification, whereas, in Model B, the aggregated classifier for associated features only had 'other' in its surrounding. Thus, the context resulted in better prediction of some classes in Model A. Moreover, this observation was mainly found in positive finding sub classes where there is more variability in the description of the findings, for example, classes like margin, morphology, distribution and associated features at third level under positive finding.

In the hierarchical models, there is not much misclassification with the global (first level) other class but with sub level other classes belonging to the same higher level, e.g., positive finding/calcification/distribution gets misclassified as positive finding/calcification/other. From this, we can conclude that good quality reports (having non-other classes) may be predicted to be of poor quality (having only other classes) but no poor quality report will be predicted to be of good quality. So, for the purpose of quality assurance, our aim to identify the poor quality reports can be solved by our models. Table 6 provides an overview of the number of reports containing each class. The shape and the density class of mass existed in only 3 reports and 1 report, respectively out of 108 mammography reports. These were the two classes reported least in the findings (a similar finding was also reported by Houssami et al. [22] in 2004). Whether this was a mistake in reporting or the observation from the images were not important enough to be reported, cannot be said. Another point is that 97 out of 108 reports contained breast composition, which are more than reported by Houssami et al. [22] (24 percent). According to ACR BI-RADS guidelines, all reports should contain breast composition, thus this type of analysis can be extended for quality assurance.

Through the automatic structuring models developed in this work, the information in the reports can be harvested and used for other purposes, for example, to generate overview statistics (e.g., "how many patients had lesions in their right breast?"). It will also support referring clinicians to read the report and gather the important information very quickly. The referring clinicians can be given a standardized semi-structured visualization of the reports and more importantly, the radiologists will not have to change their style of writing for making the reports more readable.

The most similar work to ours is Esuli et al. [18], for information extraction from mammography findings in Italian, but they had only 9 classes. Their annotation structure was not hierarchical, but they used cascaded, two-stage CRF for building their model. They had 500 labeled mammography reports (which is a lot more than what we have) and they achieved better F_1 score (0.873) than our model on these 9 classes. In another work, Hassanpour and Langotz [16] applied CRF for information extraction in chest CT radiology reports written in English. They had more number of reports and less classes compared to ours, i.e., 150 reports and 5 classes and achieved an F_1 score of 85.3 percent. We can say that although the F_1 score of our models (0.78) are not as good as the above models (0.87 & 0.85), we predict a far greater number of classes (34 classes), with much less training data (108). Increasing training data might increase the performance of our models as well. We expect our model to perform similarly for radiology reports written in languages similar to Dutch, e.g., German and English. Our models can also be applied to radiology reports for other medical conditions, e.g., ultrasound for breast cancer, chest CT, by adapting the model to their respective reporting structures.

8 Conclusion and Future Work

We developed a method for automatic structuring of Dutch free-text radiology reports on breast cancer for quality assurance. We follow a post-structuring paradigm: structuring is performed after the radiologists have written the report in free-text format.

We have addressed three tasks on breast cancer radiology reports: heading identification, content identification and automatic structuring using the BI-RADS standard. Heading and content were identified with an F_1^M score of 0.97 and 0.94 respectively using SVM. For automatic structuring, hierarchical CRF (F_1^M = 0.78) performed better than baseline CRF (F_1^M = 0.71), while Model A and B did not show any significant difference.

From the point of view of quality assurance, heading and content contribute to identification of the presence of indication of examination, findings and conclusion. A post-processing step can be performed to check if the content corresponds to the correct heading. Automatic structuring is used to check the presence of clear description of findings. According to BI-RADS, findings should contain mass, calcification, asymmetry, architectural distortion and associated features. Our model structures the findings automatically into these concepts, further generating a semi-structured XML format. This provides a platform to check the presence of important concepts. Another important information that must be present in reports is breast composition. Our model predicts breast composition with $F_1 = 0.94$.

As future work, the presence and quality of BI-RADS category can be evaluated. Based on findings, BI-RADS category can be predicted to check how well it was assigned. More reports can be labeled to get more training data. Development of a prototype and actual trial in clinical practice can be done. One of the limitations of our work is that the findings from the mammography study were manually extracted from the radiology report. A future work can be to train the system to recognize mammography, ultrasound and MRI findings and then use the mammography findings for automatic structuring. Another limitation is due to predictions occurring at 3 levels in our model, our model has the problem of error propagation. If the first level classifiers make an error, it gets propagated to the next level and makes the prediction of second and third level classifiers wrong. Our models do not contain a way to mitigate the error propagation. One way to handle this can be use of factorial CRF which jointly predicts classes at all the levels.

References

[1] R. Shikhman and L. K. Ana, "Breast, Imaging, Reporting and Data System (BI RADS)." StatPearls [Internet]. *StatPearls Publishing*, 2017.

[2] H. H. Abujudeh, R. Kaewlai, B. A. Asfaw, and J. H. Thrall, "Quality initiatives: Key performance indicators for measuring and improving radiology department performance," *Radiograph.*, vol. 30, no. 3, pp. 571–580, 2010.

[3] A. J. Johnson, J. Ying, J. S. Swan, L. S. Williams, K. E. Applegate, and B. Littenberg, "Improving the quality of radiology reporting: A physician survey to define the target," *J. Amer. College Radiol.*, vol. 1, no. 7, pp. 497–505, 2004.
[4] C. E. Kahn Jr, C. P. Langlotz, E. S. Burnside, J. A. Carrino, D. S. Channin, D. M. Hovsepian, and D. L. Rubin, "Toward best practices in radiology reporting," *Radiol.*, vol. 252, no. 3, pp. 852–856, 2009.
[5] N. Strickland, "Quality assurance in radiology: Peer review and peer feedback," *Clinical Radiol.*, vol. 70, no. 11, pp. 1158–1164, 2015.
[6] C. Yang, C. J. Kasales, T. Ouyang, C. M. Peterson, N. I. Sarwani, R. Tappouni, and M. Bruno, "A succinct rating scale for radiology report quality," *SAGE Open Med.*, vol. 2, 2014, Art. no. 2050312114563101.
[7] C. L. Sistrom and C. P. Langlotz, "A framework for improving radiology reporting," *J. Amer. College Radiol.*, vol. 2, no. 2, pp. 159–167, 2005.
[8] L. H. Schwartz, D. M. Panicek, A. R. Berk, Y. Li, and H. Hricak, "Improving communication of diagnostic radiology findings through structured reporting," *Radiol.*, vol. 260, no. 1, pp. 174–181, 2011.
[9] A. J. Johnson, M. Y. Chen, J. S. Swan, K. E. Applegate, and B. Littenberg, "Cohort study of structured reporting compared with conventional dictation," *Radiol.*, vol. 253, no. 1, pp. 74–80, 2009.
[10] S. Pathak, J. van Rossen, O. Vijlbrief, J. Geerdink, C. Seifert, and M. van Keulen, "Automatic structuring of breast cancer radiology reports for quality assurance," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2018, pp. 732–739.
[11] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," *Radiograph.*, vol. 26, pp. 1595–1597, 2006.
[12] C. L. Sistrom and J. Honeyman-Buck, "Free text versus structured format: Information transfer efficiency of radiology reports," *Amer. J. Roentgenology*, vol. 185, no. 3, pp. 804–812, 2005.
[13] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, and D. Page, "Information extraction for clinical data mining: A mammography case study," in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2009, pp. 37–42.
[14] D. A. Sippo, G. I. Warden, K. P. Andriole, R. Lacson, I. Ikuta, R. L. Birdwell, and R. Khorasani, "Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing," *J. Digit. Imag.*, vol. 26, no. 5, pp. 989–994, 2013.
[15] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.
[16] S. Hassanpour and C. P. Langlotz, "Information extraction from multi-institutional radiology reports," *Artif. Intell. Med.*, vol. 66, pp. 29–39, 2016.
[17] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 580–587, 2011.
[18] A. Esuli, D. Marcheggiani, and F. Sebastiani, "An enhanced CRFs-based system for information extraction from radiology reports," *J. Biomed. Inform.*, vol. 46, no. 3, pp. 425–435, 2013.
[19] R. K. Taira, S. G. Soderland, and R. M. Jakobovits, "Automatic structuring of radiology free-text reports," *Radiograph.*, vol. 21, no. 1, pp. 237–245, 2001.
[20] E. A. Sickles, C. J. D'Orsi, L. W. Bassett, et al., "ACR BI-RADS® Mammography," in *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology, 2013, https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads/Permissions
[21] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural Language Processing Using Very Large Corpora*. Berlin, Germany: Springer, 1999, pp. 157–176.
[22] N. Houssami, J. Boyages, K. Stuart, and M. Brennan, "Quality of breast imaging reports falls short of recommended standards," *Breast*, vol. 16, no. 3, pp. 271–279, 2007.

Shreyasi Pathak received the BTech degree in information technology from the National Institute of Technology, Durgapur, India, in 2016, and the MSc degree in computer science (specialization in data science) from the University of Twente, the Netherlands, in 2018. She is currently working as a junior researcher with the University of Twente. Her research interests include data mining, information extraction, natural language processing, machine learning, and biomedical data analysis.

Jorit van Rossen received the BSc degree from the University College Utrecht (International Honors College of Utrecht University, Liberal Arts and Sciences), in 2009, and the MSc/MD degrees from Utrecht University, in 2013. She is currently working as a radiology resident with ZGT Almelo and the University Medical Center Groningen. Her research interests include natural language processing and machine learning in relation to the medical and specifically radiological field.

Onno Vijlbrief received the medical degree from Leiden University, the Netherlands. After this, he finished his radiology residency and fellowship training in neuro- and head and neck radiology with the Hague Medical Centre and the Leiden University Medical Centre. He currently works as a neuro- and head and neck radiologist for MRON and ZGT. His research interests are focused on using healthcare informatics to improve hospital processes, clinical decision making, visual interpretation, and improving patient outcome through the use of text and process mining, natural language processing, and information extraction from electronic health records and hospital image archives.

Jeroen Geerdink received the BSc degree in electrical engineering from the Saxion University for Applied Science, in 1991. He is currently an innovation manager with the Hospital Group Twente, the Netherlands. His fields of interests include data mining, machine learning, system interoperability, and imaging informatics.

Christin Seifert received the PhD degree from the University of Graz, Austria, in 2012. She is currently an assistant professor with the Faculty of EEMCS, University of Twente. Her publication list comprises more than 100 peer-reviewed publications in the fields of data mining, natural language processing, and information visualization. Her core research interests include explainable machine learning, and data mining, as well as intersections with human-computer interaction, information visualization, and information retrieval.

Maurice van Keulen received the MSc and PhD degrees from the University of Twente, in 1992 and 1997, respectively. He is currently an associate professor with the Faculty of EEMCS, University of Twente. He has published more than 120 research papers in various journals, conferences, and workshops. His research interests include data integration, data quality, data interoperability, data cleaning, probabilistic databases, natural language processing, machine learning, and database systems.