UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance

Shreyasi Pathak

Masters in Computer Science
Specialization: Data Science and Smart Services

Master Thesis
27th August, 2018

External Supervisors:
drs. Onno Vijlbrief
Email: o.vijlbrief@zgt.nl
Jeroen Geerdink
Email: J.Geerdink@zgt.nl
drs. Jorit van Rossen
Email: j.vrossen@zgt.nl

Radiology Department
Ziekenhuis Groep Twente (ZGT)
Geerdinksweg 141
7555 DL Hengelo
The Netherlands

Supervisors:
Dr. Ir. Maurice van Keulen
Email: m.vankeulen@utwente.nl
Dr. Christin Seifert
Email: c.seifert@utwente.nl

Data Management and Biometrics
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE

MASTER THESIS

Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance

Author:
Shreyasi PATHAK

Supervisors:
Dr. Ir. Maurice VAN KEULEN
Dr. Christin SEIFERT

External Supervisors (ZGT):
drs. Onno VIJLBRIEF
Jeroen GEERDINK
drs. Jorit VAN ROSSEN

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in

Computer Science
Data Science and Smart Services
Datamanagement and Biometrics Research Group
Faculty of Electrical Engineering, Mathematics and Computer Science

August 27, 2018

Declaration of Authorship

I, Shreyasi PATHAK, declare that this thesis titled, "Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Shreyasi Pathak

Date: 27.08.2018

Abstract

Hospitals often set protocols based on well defined standards to maintain 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 of the report, ii) check whether the information under each section corresponds to the section heading, iii) convert the free-text detailed findings in the report to a semi-structured format. Top level structure and content of report were predicted with an $\mathbf{F}_1$ score of 0.97 and 0.94 respectively using Support Vector Machine (SVM). For automatic structuring, our proposed hierarchical Conditional Random Field (CRF) outperformed the baseline CRF with an $\mathbf{F}_1$ score of 0.78 vs 0.71. The third step generates a 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.

Acknowledgements

The past 8 months that I have been working on my master's thesis, have been an amazing experience for me. I learnt the skills to conduct research, write a research paper and how to work collaboratively. This would not have been possible without my supervisors. I would like to thank Maurice van Keulen and Christin Seifert for being so supportive and helpful throughout the project, for all the brain-storming discussions and critical feedback. Every time I had meetings with both of you, I would always feel very motivated and encouraged. Maurice, thank you for always creating a visualization out of the complex things and making it simpler. Christin, thank you for all your detailed feedback, for sending me helpful websites and sharing your books with me, so that I could understand something better. I would like to thank Jorit van Rossen and Onno Vijlbrief for taking time to explain me the related knowledge from the medical domain and for labeling the reports. Thank you for patiently answering my numerous questions and emails. I would like to thank Jeroen Geerdink for setting up the system for my work, for the dataset, for always helping me out with difficulties related to the hospital and for giving me an amazing overview of the project on the very first day. It was a great pleasure working under all of your supervision and I learnt a lot from all of you.

I would like to extend special thanks to my parents, my boyfriend, and my brothers for always being there through this roller coaster ride. It would not have been possible without your support. During the times that I would feel down, you were always there to listen and motivate me through numerous video calls.

I would also like to thank my friends at the university, who helped me survive these two tough years of masters, with some amazing get-together. Finally, I would like to thank you, the reader, for taking time to read my thesis.

Preface

This master thesis is divided into two parts. The first part consists of the research paper on my master's project, containing a concise overview of the work and the important results. This research paper was a deliverable for the masters research honours programme that I participated in and this paper was also submitted to a workshop in a conference. The second part consists of a detailed appendix explaining the things that could not be in the paper like an elaborate motivation and literature review, explanation of the models and more results.

Contents

List of Figures

List of Tables

Part I

Research Paper

Automatic Structuring of Breast Cancer Radiology Reports for Quality Assurance

Shreyasi Pathak
University of Twente
Enschede, Netherlands
s.pathak@student.utwente.nl

Jorit van Rossen
Hospital Group Twente (ZGT)
Hengelo, Netherlands
j.vrossen@zgt.nl

Onno Vijlbrief
Hospital Group Twente (ZGT)
Hengelo, Netherlands
o.vijlbrief@zgt.nl

Jeroen Geerdink
Hospital Group Twente (ZGT)
Hengelo, Netherlands
J.Geerdink@zgt.nl

Christin Seifert
University of Twente
Enschede, Netherlands
c.seifert@utwente.nl

Maurice van Keulen
University of Twente
Enschede, Netherlands
m.vankeulen@utwente.nl

***Abstract*—Hospitals often set protocols based on well defined standards to maintain 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 of the report, ii) check whether the information under each section corresponds to the section heading, iii) convert the free-text detailed findings in the report to a semi-structured format. Top level structure and content of report were predicted with an F_1 score of 0.97 and 0.94 respectively using Support Vector Machine (SVM). For automatic structuring, our proposed hierarchical Conditional Random Field (CRF) outperformed the baseline CRF with an F_1 score of 0.78 vs 0.71. The third step generates a 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, Radiology Reports, Conditional Random Field**

I. Introduction

Medical reports are essential for communicating the findings of imaging procedures with referring physicians, who further treat the patients by considering these reports. Thus, medical reports are very important for diagnosis of diseases, which brings forward the need of their quality assurance.

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], which is a classification system proposed by American College of Radiology (ACR), to represent the malignancy risk of breast cancer of the 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 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. Another method can be 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.

Thus, 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 quality of the report by how well the reports conform to the reporting standards as set by ACR BIRADS. Concretely, we (i) identify the top-level structure from the reports (henceforth, referred to as heading identification), (ii) verify if the report contents are placed under the correct top-level headings (referred to as content identification), and, (iii) automatically convert the free-text 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 further use, we generate a semi-structured XML format for the automatic structuring (Table I). We focus on Dutch radiology reports on breast cancer; for automatic structuring we focus on findings from mammography imaging modality.

In the remainder of this paper, we first review structured reporting initiatives and application of natural language processing to radiology reports (Section II). Section III describes the dataset. Our approach to heading and content identification, and automatic structuring is detailed in Section IV. We describe our experimental setup in Section V followed by experimental results in Section VI. We discuss the implication of our results and some future work in Section VII.

II. Related Work

In this section, we will discuss structuring initiatives for radiology reporting, followed by various natural language processing techniques applied in radiology.

A. Structured Reporting Initiatives

Accuracy, clarity, timeliness, readability, 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 through out the world. Breast Imaging Reporting and Data System (BI-RADS) is a very successful attempt by ACR at standardizing the language for breast cancer reporting [1]. RadLex [10] 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. They point out that it is 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 [11] tested information extraction from free-text and structured reports. It was 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, Panicek, Berk, Li and Hricak [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. Whereas, a study by Johnson, Chen, Swan, Appelgate and Littenberg [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, past works have shown that firstly, structured reporting and standard language are important for quality of report. But structured reporting should be such that it should not impose restriction on the radiologist. Secondly, structuring reporting can help in addressing quality assurance.

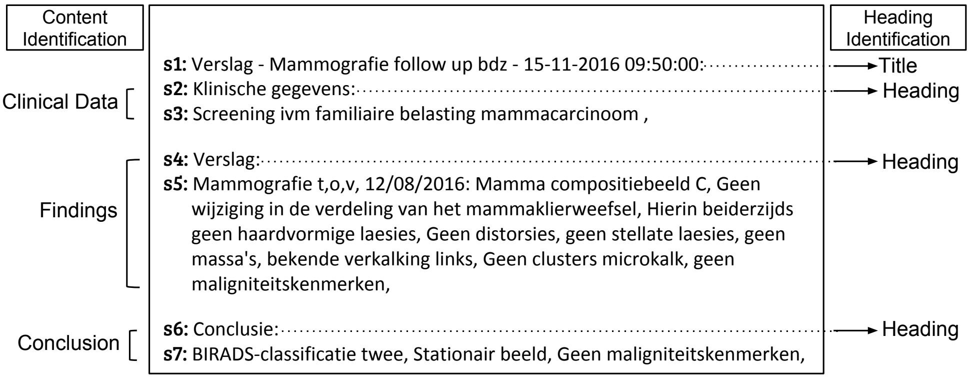
B. Natural Language Processing in Radiology

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-based* 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 tasks in radiology reports on breast cancer. Nassif et al. [12] developed a rule-based system in 2009 to extract BI-RAD related features from a mammography study. The system was tested on 100 radiology reports manually tagged by radiologists, resulting in a precision of 97.7% and a recall of 95.5%. Sippo et al. [13] developed a rule-based NLP system in 2013 to extract the BI-RAD 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% and a precision of 96.6%.

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)* [14] are some of the ML approaches used for sequence labeling. Hassanpour and Langlotz [15] 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 [16] 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 [17] 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-based approaches. For example, Taira, Sodrland and Jakobovits [18] developed a automatic structuring of free-text thoracic radiology reports using some rule-based and some statistical and machine learning methods like maximum



s1: Verslag - Mammografie follow up bdz - 15-11-2016 09:50:00:
s2: Klinische gegevens:
s3: Screening ivm familiaire belasting mammacarcinoom ,
s4: Verslag:
s5: 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,
s6: Conclusie:
s7: BIRADS-classificatie twee, Stationair beeld, Geen maligniteitskenmerken,

Fig. 1: Example of a breast cancer radiology report

entropy classifier. We want to develop a fully automated system without any rule creation involved from experts, which is why we will not follow hybrid approach.

In this work, we apply machine learning-based approaches to avoid manual rule construction and use CRFs which have been shown to provide high performance on sequence labeling.

III. Corpus: Radiology Reports on Breast Cancer

According to BI-RADS [19], 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 consisting 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 a hospital in The Netherlands. The reports were produced by dictation from trainee or consultant radiologist, into an automatic speech recognition system. These automatically generated reports are further cross-checked with the dictation, by radiologists or secretary. The reports are anonymized such that they do not contain patient identity data like patient id, name, data of birth and address. A sample report is shown in Fig. 1. 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) (see Fig. 1). 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 breast cancer 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 a sentence means start of a new line, resulting in 1591 sentences in total. In Fig. 1, 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 of reports, 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, 45 reports by the trainee and 47 by the consultant. After labeling, these 45 reports and 47 reports were analyzed to highlight any inter-annotator discrepancy, which were further resolved by the annotators.

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 IV-C). 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

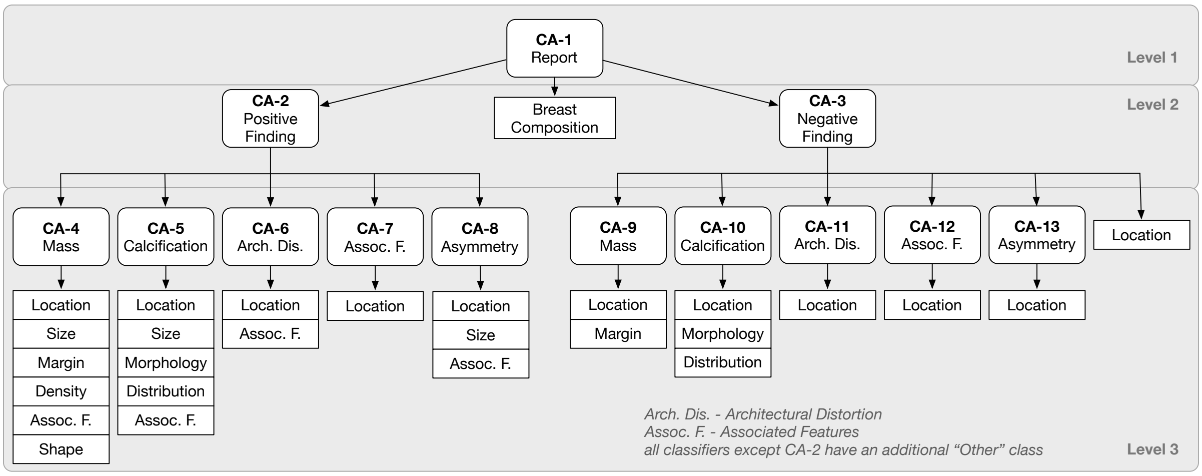


Fig. 2: 3-level annotation scheme for automatic structuring of mammography findings
(Hierarchical Conditional Random Field Model A (Section IV-C2))

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. 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.

IV. Approach

In this section, we describe our approach for the three sub-goals – heading identification, content identification, and automatic structuring of findings from mammography study.

A. Heading Identification

a) Feature extraction: Reports were separated into sentences as explained in Section III. 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% 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.

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 length of the sentence and the symbol at the end of sentence were also tested as features but they did not improve performance and were not considered further.

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

B. 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 IV-A.

C. 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 I, where the first column shows a free-text mammography finding report and the second column shows the semi-structured XML version. Let **X** be a mammography finding report, consisting of a sequence of tokens, **x**=(x_1,x_2,..x_t,..,x_n) and the task is to determine a corresponding sequence of labels **y**= (y_1,y_2,..y_t,..,y_n) for **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, meaning 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] [14],

[1] Classifiers were built using Python scikit-learn package

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

TABLE I: Example of structuring of free-text mammography finding

Free-text Report	Structured Report
Mammografie t,o,v, 22/09/2016: 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 kalk, 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⟩

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.

a) 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, **x**, which were transformed according to the IOB tagging scheme [20]. 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 I, *"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.

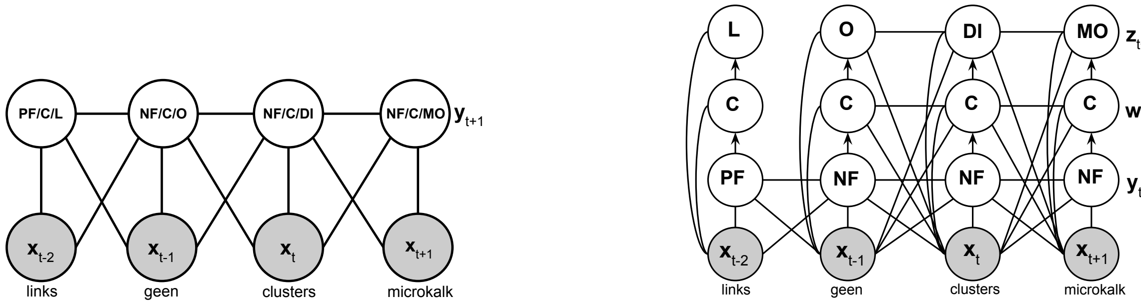
b) Feature Extraction: Each extracted token, x_t, is represented by a feature vector $\mathbf{x}_t$ for LC-CRF, including linguistic features of the current token, x_t. and also features of the previous token, x_{t-1}, and the next token, x_{t+1}. A feature vector $\mathbf{x}_t$ consists of the following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

Below, we describe the 3 models for automatic structuring:

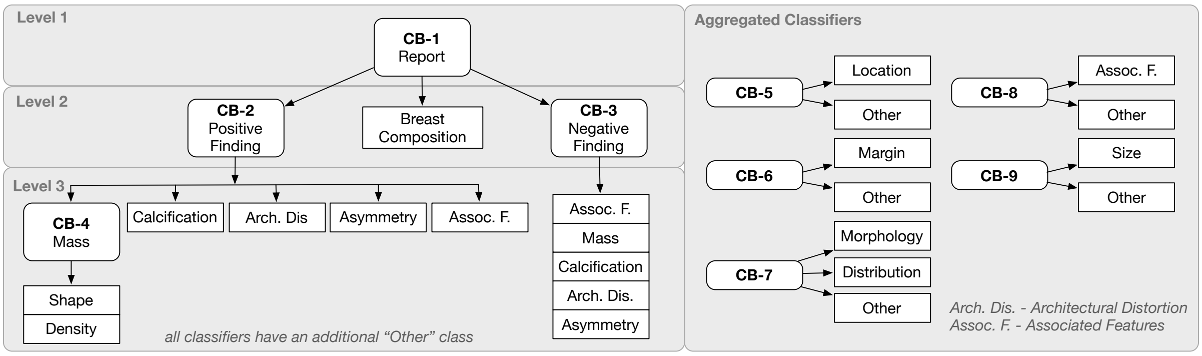
1) Baseline Model: As baseline, we used one LC-CRF classifier, as described at the starting of Section IV-C, to predict the complete label (concatenation of labels at the 3 levels) of a token and as input to the classifier, we used the feature vectors described in *Feature Extraction* (Section IV-Cb). For example, the LC-CRF classifier will predict the tokens *clusters* and *microkalk* as *NF/C/DI* and *NF/C/MO* respectively (see Table I). 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 $\mathbf{x}_t$, $\mathbf{x}_{t-1}$, $\mathbf{x}_{t+1}$ 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.

2) 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



(a) (b)

Fig. 3: Graphical representation of a) baseline CRF model and b) hierarchical CRF model, for input feature vectors $\mathbf{x}_{t-2}$ to $\mathbf{x}_{t+1}$={links geen clusters microkalk}



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

Fig. 4: Hierarchical Conditional Random Field Model B

further classifying the tokens predicted as positive finding by CA-1, another (CA-3) 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 NP/C/DI and NP/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.

3) Hierarchical CRF with Combined Classes: As can be seen 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, then the class with the highest a-posteriori probability is selected.

V. 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 c_1 is the harmonic mean of precision and recall of that class and is defined as

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

with TP being the number of true positives, FP - false positives and FN - false negatives. As our problem is a multiclass problem, the TP, FN, FP of a class are calculated according to one-vs-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 our classifiers at 3 levels: i) token-level (TL), ii) partial phrase-level (PP), and iii) complete phrase-level (CP). At the token-level, we consider all the token labels in the dataset to calculate the TP, TN, FP, FN scores of a class. At the partial phrase-level and the complete phrase-level, 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

TABLE II: 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

and match with the corresponding predicted labels. Phrase-level scores are important from the radiologists' point of view. They care about how well their phrases are matching. Table IIIa 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 $2 * 2/(3+3) = 0.66 > 0.65$, 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 matching (TL F_1).

For heading and content identification, we evaluated NB, SVM and RF models, using 5-fold cross validation on 180 reports. 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 IIc). '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. All codes associated with this paper are available as open source[3].

VI. Results

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

A. Heading and Content Identification

Table IIa shows that headings were identified with a F_1^M score of 0.96 both by SVM and NB and sentences which were not headings were identified with a F_1^M score of 0.98 by SVM. For both heading and not heading classes, SVM and NB performed better than RF. For title class, RF performed better. Table IIb shows that the SVM performed better for predicting the classes conclusion, clinical data, title and findings with a F_1^M scores of 0.92, 0.94, 0.99 and 0.94 respectively.

B. Automatic Structuring

Table IIc compares the performance of our LC-CRF baseline model to the hierarchical LC-CRF Models A and B. Both, Model A and B (F_1^M=0.78) outperformed the baseline model (F_1^M=0.71). No difference in performance was observed within Model A and B. Without the not important *other (O)* class, the model B has a F_1^M of 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 less than 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 IIIb 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 better than complete phrase-level measure. Partial phrase accuracy (PP-Acc) is at least as good as complete phrase accuracy (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.

Table IV shows the performance obtained for the some of the global classes. 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

[3] https://www.dropbox.com/sh/y4czin4llue2t6w/AACqHRcC2pxg0zzg42JuPtQna?dl=0

TABLE III: 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

TABLE IV: F_1 measures of global classes for the 3 models of automatic structuring

Models	**BC**	**NF/AF/O**	**NF/C/O**	**NF/C/DI**	**NF/C/MO**	**NF/MS/O**	**NF/MS/MA**	**PF/C/O**	**PF/C/SI**	**PF/C/L**	**PF/MS/L**	**PF/MS/MA**	**PF/C/AF**	**PF/AS/O**
Baseline	0.89	**0.96**	**0.81**	0.98	0.95	**0.93**	**1.00**	**0.45**	0.00	0.50	0.30	0.53	0.00	0.00
Model A	**0.94**	**0.96**	0.76	0.98	0.91	0.88	0.96	0.37	0.00	0.44	0.40	**0.72**	**0.18**	**0.58**
Model B	**0.94**	**0.96**	**0.81**	**0.99**	**0.97**	0.89	0.97	0.37	**0.22**	**0.60**	**0.47**	0.70	0.00	0.56
#Instances	622	397	148	54	56	210	35	138	14	68	139	59	33	172

TABLE V: Error propagation through classifiers at the 3 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

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.

Table V gives an indication on 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. 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.

VII. Conclusion and Future Work

We have addressed three tasks for the purpose of quality assurance of radiology reports: heading identification, content identification and automatic structuring using BIRADS standard. Heading and content were identified with a 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 0.94 F_1 score.

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. The approach taken in this research can also be extended to reports for other conditions, written in other languages.

References

[1] *Breast imaging reporting and data system.* BI-RADS Committee, American College of Radiology, 1998.

[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,” *Radiographics*, 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,” *Journal of the American College of Radiology*, 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,” *Radiology*, vol. 252, no. 3, pp. 852–856, 2009.

[5] N. Strickland, “Quality assurance in radiology: peer review and peer feedback,” *Clinical radiology*, 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 medicine*, vol. 2, p. 2050312114563101, 2014.

[7] C. L. Sistrom and C. P. Langlotz, “A framework for improving radiology reporting,” *Journal of the American College of Radiology*, 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,” *Radiology*, 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,” *Radiology*, vol. 253, no. 1, pp. 74–80, 2009.

[10] C. P. Langlotz, “Radlex: a new method for indexing online educational materials,” 2006.

[11] C. L. Sistrom and J. Honeyman-Buck, “Free text versus structured format: information transfer efficiency of radiology reports,” *American Journal of Roentgenology*, vol. 185, no. 3, pp. 804–812, 2005.

[12] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, and D. Page, “Information extraction for clinical data mining: a mammography case study,” in *Data Mining Workshops, 2009. ICDMW’09. IEEE International Conference on.* IEEE, 2009, pp. 37–42.

[13] 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,” *Journal of digital imaging*, vol. 26, no. 5, pp. 989–994, 2013.

[14] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” 2001.

[15] S. Hassanpour and C. P. Langlotz, “Information extraction from multi-institutional radiology reports,” *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 2016.

[16] M. Torii, K. Wagholikar, and H. Liu, “Using machine learning for concept extraction on clinical documents from multiple data sources,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 580–587, 2011.

[17] A. Esuli, D. Marcheggiani, and F. Sebastiani, “An enhanced crfs-based system for information extraction from radiology reports,” *Journal of biomedical informatics*, vol. 46, no. 3, pp. 425–435, 2013.

[18] R. K. Taira, S. G. Soderland, and R. M. Jakobovits, “Automatic structuring of radiology free-text reports,” *Radiographics*, vol. 21, no. 1, pp. 237–245, 2001.

[19] E. A. Sickles, C. J. D’Orsi, L. W. Bassett, and et al, *ACR BI-RADS Mammography. In*, Reston, VA, 2013.

[20] L. A. Ramshaw and M. P. Marcus, “Text chunking using transformation-based learning,” in *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.

Part II

Master Thesis

Chapter 1

Introduction

In this section, we give an overview of the field of radiology and problems associated with the current way of radiology reporting. Then, we introduce our approach to address the problems and explain the research questions associated with it.

1.1 General Overview of Radiology

Radiology is the science of diagnosing diseases using medical images. There are different imaging techniques like X-ray radiography, ultrasound, Magnetic resonance imaging (MRI). The radiographer is assigned with the task of acquiring medical images through these techniques. The radiologist interprets the images produced by the radiographer and writes a report listing his findings and diagnosis. The report is then sent to the referring physician who diagnoses and treats patients by considering these radiology reports.

1.2 Current Scenario and Problems Associated with it

The radiology reports are written by the radiologists currently in free-text format. The contents of the report are written in a narrative style in the order and format as deemed fit by the radiologist. Thus, absence of standardized structure in the reports creates several problems, as listed below:

1. Difficulty in information extraction by radiologists as well as physicians.
2. Different writing styles of different radiologists makes readability of reports hard for physicians.
3. Hard to assess quality of radiology reports and analyze how well the radiologists are conforming to the standards.

The problems listed above will be explained in detail in the following paragraphs.

The first problem listed above is difficulty in information extraction. Often, the radiologists or the referring physicians need to find answer to questions, such as "What initial symptoms were troubling patient X which required medical imaging?". To answer this kind of question, the report has to be scanned through to find where the initial symptoms have been listed. A more difficult and elaborate question can be, "How many patients were diagnosed by breast tumor in their lower outer quadrant of their right breast?". To answer this, the reports of all the patients need to be analyzed manually. According to radiologists, finding answer to these type of questions get very difficult if the information in the reports are unstructured.

The second problem is free-text writing styles decreases readability of reports. As there is no strict structured format for writing the reports, the radiologists write the report in their own structure. For example, in general, the indication/clinical data part of the report are written at the beginning of the report. It may be that some radiologists write the clinical data at the end of the report or may be they forgot to write it at the beginning. This creates difficulty for the referring physician as he has to search through the whole report to find the information he needs and adapt himself to different writing styles of different radiologists. This results in decrease of readability, which further gives rise to reports not being read intensively and some of the information in the reports remaining unused.

The third problem is free-text writing style makes it hard to assess quality of the reports written by the radiologists. Among various things, quality of radiology reports can be assessed by its structure, standardized use of language and how well the radiologists are adhering to set reporting standards. For example, Breast cancer radiology reports use "Breast imaging-reporting and data system"(BI-RAD), which is a classification system proposed by American College of Radiology, to represent the malignancy risk of breast cancer of the patient. It was implemented to standardize risk assessment and quality control for mammography and to provide a general understanding of the findings to non radiologists. BI-RAD lexicon lists the terms that can be used to report different findings after analyzing medical images. Based on these findings, the radiologist gives a BI-RAD assessment category (can be numbered from 0-6) at the conclusion of the report. Often the radiologists want to know how correctly they are doing their job of assigning a BI-RAD score to a report. For quality assurance purpose, it would be interesting to perform a check on whether the BI-RAD assessment by the radiologists correspond to the prescribed standard. Another important quality check is if the reports are following BI-RADS guidelines. A study [13] done on 244 breast cancer imaging reports from 2004 reported that only half of the reports were concordant with reporting standards. The least reported variables were breast density (reported in 24% reports), lesion depth (37%), lesion shape (55% for mammography) and location (59%). With the current free-text format, it is hard to analyze if all the necessary information are mentioned in the report and how well the BI-RAD score given by the radiologist, corresponds with the findings. This makes the quality assurance difficult in free-text format reports.

1.3 Possible Solutions and Our Approach

Out of the three problems discussed in Section 1.2, we decided to focus our work on addressing the third problem - Quality assurance. According to literatures [16, 17], quality of a radiology report can be assessed by its accuracy, timeliness, clarity, organization etc. For our project, quality assurance means how well the report conforms with the well-defined protocols of radiology report writing and this in turn will help in checking clarity, accuracy and organization. At the basis of this quality assurance, we will be converting the free-text reports to a structured format for making the task of comparison of reports to well-defined protocols easier.

Free-text reports can be converted to structured format using any of the following two approaches. One of them is a pre-defined report structure provided to the radiologists and asking them to write the reports according to it. This can be realized by developing a report structuring software which guides the radiologist during writing, thus restricting his style and format of writing. An already existing work [15] shows that this type of system resulted in lower quality reports. Thus, this type of guiding systems deteriorates the task of radiologists in interpreting the images.

Another approach is converting the free-text reports to structured format after the reports have been written by the radiologists in their own style. This method does not impose any particular structure of writing on the radiologists. A system can be developed which takes free-text reports written by the radiologist and convert it to a structured format automatically without involvement from the radiologist. Thus, the radiologists can concentrate more on the task of interpreting images and listing the findings rather than thinking how it should be written, which helps in maintaining accuracy in the task of image interpretation.

For this project, we decided to adopt the second approach so as not to decrease radiologist's performance. To summarize, our aim is to develop a system which will automatically structure the radiology reports for the purpose of quality assurance and for our project, we focus on breast cancer reports. We will convert the reports into a semi-structured format and not a table-of-contents structured format. The difference between these two are that in semi-structured format, the information present in the report are labeled and structured. Whereas, in table-of-contents structured format, a table is constructed having entries for all possible information that can be in a report. For each report, only those cells are filled up, corresponding to which there is information in the report and other cells remain empty. Quality of a breast cancer radiology report will be assessed according to ACR BI-RADS rules [30]. The department of radiology at Hospital Group Twente (ZGT) provided with the breast cancer radiology reports.

1.4 Research Questions

The main research question of our research is as follows:

"To what degree can we successfully conduct quality assurance of radiology reports using machine learning algorithms?"

The main research question is divided into the following sub-research questions:

1. (RQ-1) *How can we identify the most apparent top level structure from the report using machine learning?*
   The ACR BIRADS [30] mention that a report should contain indication of examination (clinical data), clear description of findings and a conclusion. We performed heading identification to identify the top-level structure.

2. (RQ-2) *How can we automatically verify if the information in the report has been placed under the correct top level sections (from RQ-1)?*
   It was further be checked if the headings identified in RQ-1 contain the information corresponding to the heading. This is needed, as it may sometimes happen that the heading of clinical data is there but the content under it is of findings. Thus, a check was done for the presence of mainly 3 sections – clinical data, findings and conclusion.

3. (RQ-3) *To what extent can we automatically convert the free-text findings from the report into a detailed structured format?*
   We took the findings section from the report and converted it to a structured format for checking if the findings contain all the necessary information. The findings section can have findings from different imaging modalities like mammography, ultrasound and MRI. For this project, we only considered findings from mammography. In consultation with the radiologists, a structure of the mammography findings was developed based on ACR BI-RADS. Mammography finding contains mass, asymmetry, calcification, architectural distortion and associated features and our structure was created to identify these classes.

The rest of the thesis is organized as follows: Chapter 2 presents a elaborate literature review about the work done on quality of radiology reports, structuring initiatives taken and natural language techniques applied in radiology. Chapter 3 describes the theoretical background of the machine learning algorithms and necessary knowledge from domain of radiology reports. Chapter 4 presents the approach taken to solve the research questions. As, most of it was already described in paper, it only contains those parts that could not be written in the paper. The dataset has already in explained in the paper and thus, part 2 of the thesis does not contain it. Chapter 5 has extra experiments that could not be in the paper and their implication. Chapter 6 provides with a short conclusion discussing the research questions.

Chapter 2

Related Work

This section contains discussion of various literatures on evolution of radiology reporting and how the radiology reports can be analyzed by computer to extract meaningful data out of it. Section 2.1 talks about the history of radiology reporting and the expected quality of the reports. In Section 2.2, we discuss the structured report initiatives that have been taken till now and the viewpoint of the two reporting style – free-text and structured. Section 2.3 talks about different methods developed for automatic analysis of radiology reports. It introduces natural language processing and its two types of techniques used for processing the data – rule based and machine learning based. Literatures using different types of machine learning approaches for processing radiology reports have been discussed. There is also a short discussion on deep learning techniques being used for processing radiology reports.

2.1 Introducing Radiology Reporting and its Qualities

Radiology reports are very essential for communicating the findings of imaging procedure with referring clinicians and patients. Based on these reports, the referring clinician gets a better understanding of the patient's condition and decides upon the treatment. This importance of a radiology report leads to the need of their quality assurance. The reports need to be concise, clear, understandable and also need to be written correctly.

Wilhelm Rontgen, the discoverer of X-rays, published the earliest radiology report called Ueber Eine Neue Art von Strahlen [27] in 1896. The importance of good quality of radiology report was first recognized by Preston M. Hickey in 1922 [12]. Hickey wanted to assess the radiologists by looking at the quality of their radiology reports. He suggested that each radiologist interested in seeking admission to American Roentgen Ray Society (ARRS) will be required to submit 100 radiology reports with their application. He stressed on the fact that a standardized nomenclature should be used in writing radiology reports [12]. Qualities of a good radiology report was summarized by Armas [1] in 1998. He listed 6 C's namely clarity, correctness, confidence, concision, completeness and consistency as characteristics of good quality report.

In 2010, Pool and Goergen [23] did literature review to identify the important elements of a high-quality radiology report. Their aim was to identify evidences that talk about the content of radiology report and also determine the gaps among these evidences so that it can be filled up by further research. To achieve this, they reviewed 25 published papers which consisted of study methodologies such as 1 randomized controlled trial, 1 before-and-after study of interventions, 10 observational studies, 12 surveys and 1 narrative review of the literature. They also reviewed 4 guidelines of professional standards of radiology reports namely ACR, the Canadian Association of Radiologists, the Royal College of Radiologists and Society of Interventional Radiology. They found out that existing guidelines had several weakness related to scope, purpose, methods of development, stakeholder consultation etc. and there was a lot of difference in the languages used to describe images, diagnostic uncertainty. They also found that many survey participants preferred structured or itemized style of radiology report, but not many studies exist about the effect of report structure on its quality.

Another very recent paper published in 2017 talks about different ways that radiologists can make the reporting more effective by just following some simple steps [39]. It asks radiologists to organize their thoughts, be clear, take responsibility, close the loop on incidental findings, make reports readable for patients and be an expert consultant.

2.2 Structured Reporting Initiatives

There is active research going on related to improving quality of radiology report. Sistrom and Langlotz [33] identified i) standard language, ii) structured 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 through out the world. To bring standardization into effect, in 2006, Radiology Society of North America (RSNA) created a lexicon called RadLex [19] which provides standard terminology for diseases, observation and radiology procedure. Each term in RadLex also contains all its synonyms and other related terms.

It was further understood that increasing readability of radiology reports can be attained by putting the information in a *structured format*. Structured format facilitates reuse and retrieval of report content both by human readers and information systems. In the next paragraphs, we will see various literatures where the effects of structured reporting have been studied.

In a research paper [25], Reiner, Knight and Siegel talk about the evolution of radiology reports and different methods adopted in changing the free-text radiology report to structured text. They proposed a graphical system that directly maps the terminologies in the report to standardized lexicon RadLex. RSNA established a Radiology Reporting Committee to promote best practices in radiology reporting. This committee consisting of radiologists and imaging informatics experts conducted a workshop in 2008 to address the issue about structured reporting and how it should be adopted throughout radiology. The highlights of the workshop was published in 2009 by RSNA [17]. It stresses on the fact that structured reporting would help in research, teaching and clinical quality improvement. It established a framework about the contents of a radiology report such as administrative information, patient identification, clinical history, imaging technique, comparison, observations, summary or impression and signature. This paper also discusses that structured reporting can help in addressing quality assurance and subsequently lists the quality metrics that can be derived from radiology report data.

Though there had 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 [32] performed an experiment to test the accuracy and speed of the reviewers in extracting case-specific information from free-text and structured report. A web-based testing mechanism was used to give radiology reports to 16 senior medical students, who were asked to answer 10 multiple choice questions on each of the 12 cases. Students were randomly assigned either free-text or structured report. Three things were recorded while they answered the questions- the number of questions answered correctly for each case, the time taken for each case and the number of questions answered correctly per minute. At the end of the test, it was found that both the free-text and structured report resulted in similar accuracy and efficiency in information extraction. A post-experimental questionnaire was also conducted where the subjects expressed an opinion in favour of the structured report format.

In another study conducted in 2009 by Johnson and his colleagues [15], it was reported that structured reporting may be inferior to free-text reports. A cohort study of structured radiology reporting was compared with conventional dictation reports and the quality of the report was graded based on accuracy and completeness. The study involved 16 resident radiologists in the control group and 18 in the intervention group. The residents in the intervention group were asked to use commercially available structured reporting system(SRS), a point-and-click menu driven software, to create the structured format reports and the residents in the control group used the free-text dictation format. It was concluded that the structured reporting resulted in a decrease in report accuracy and completeness, which could in turn affect patient care. Accuracy decreased from 91.5 to 88.7 and completeness decreased from 68.7 to 54.3 when using SRS. The residents also complained that SRS was overly constraining and time-consuming. It did not allow them to use desired content in the report. Even then most of the residents commented that the idea of structured reporting is appealing and good reporting skills and standardized terminology are important in clinical practice.

Schwartz and his colleagues [28] did a similar study in 2011 to compare the content, clarity and clinical usefulness of conventional dictated radiology reports and template-structured reports and their outcome differed from the study conducted by Johnson [15]. Referring clinicians and radiologists found greater satisfaction with content and clarity in structured reports, but the clinical usefulness graded using a radiology report grading scale (PCOS) [26] did not differ significantly between the two types of report.

Another recent study conducted in 2015 by Powell and Silberzweig [24] showed evidence in favour of structured reporting. An online survey related to the development and experience of structured reporting was sent to the members of Association of University Radiologists and around 59.5% of the respondents reported to be satisfied with structured reports. The study showed that many radiology departments were experimenting with structured reporting and departments using structured report format had less error in their reports. But many of the radiologists stated that structured reporting had a lot of limitations like report formats are variable and complex patient images may not properly fit into this structure.

In conclusion, from the above discussion, it can be noted that literatures suggest structured reporting can help in quality assurance of reports. Literatures show that though the radiologists did not like softwares imposing structured reporting on them while writing and addressed it as *time-consuming* and *overly constraining*, they preferred structured format of reports over free-text.

2.3 Natural Language Processing in Radiology

Electronic health record (EHR), mostly in text and images format, is increasing the use of digital content and thus generating lot of new challenges and opportunities in the medical domain. These records contain a lot of information which can be used to improve clinical care. It is not possible for humans to analyze these huge amount of data and extract relevant information manually. Efficient and automated strategies are required to aid humans in making the maximum utilization of the available data and extract all relevant information. As discussed in the previous section, radiology reports, a type of EHR, are usually written in unstructured free-text format. This format of the report is not suitable for many computerized applications like automated quality assurance, clinical decision support and research. Thus, for many years, there has been a lot of research going in the field of natural language processing to automatically convert the reports to structured format so that automatic identification and extraction of information becomes easier.

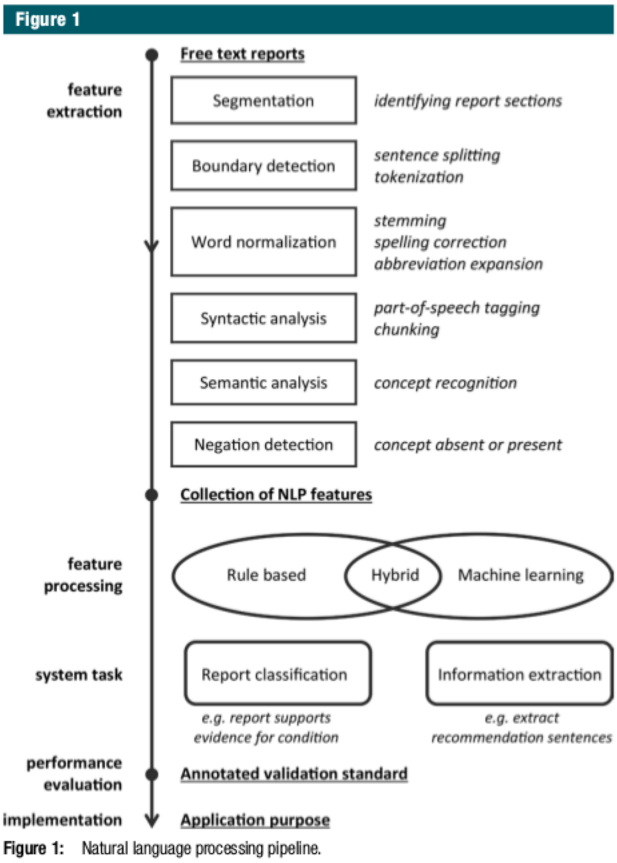


FIGURE 2.1: NLP pipeline (taken from [22])

Natural language processing (NLP) is analysis and synthesis of natural language (text and speech) and is often related to the terms text mining and information extraction as described by Hearst [11]. The use of NLP in biomedical domain and also its current state has been summarized by Demner-Fushman in 2008 [6] and in

2016 by Pons [22]. Demner-Fushman along with his colleagues reviewed the state of NLP in computerized clinical decision support(CDS) that aims at helping health care providers and public by providing easy access to information about the patient whenever needed. The authors state that though a lot of benefits and successes were observed through CDS system, widespread use of CDS systems for natural language processing research and daily practice was not observed for many years. Most of the methods were targeted at specific goals and for specific information systems. For making NLP research a success, the systems need to be easily adapted to meet new goals. The authors concluded that there is a renewed interest in NLP research in medical domain and there has been some local successes related to it. Thus, more NLP systems for CDS will develop and come into daily use. In a more recent study conducted in 2016 [22], Pons and his colleagues identified 67 relevant publication related to how radiology is benefiting from NLP. The authors summarized the methods and tools of NLP being applied into practical applications of radiology and how well they were performing.

To convert the unstructured reports to structured format, several NLP steps (shown in Fig 2.1) need to be taken which has been described well by Pons and his colleagues [22]. As shown in the figure 2.1, after the features are extracted from text, there are two types of techniques for processing these features - *i) rule based approach* and *ii) statistical approach (machine learning)*. In rule based, rules are manually created by experts to match a specific task. Dictionary based is a special case of rule based approach where a lexicon is used to match terms from the text. In statistical approach, machine learning algorithms are trained to automatically recognize pattern. This statistical approach is also called statistical machine learning approach, but we will refer this as machine learning in rest of our thesis. When rule based and machine learning are combined together for processing the features it is called hybrid approach. In this approach, rules are generally used to improve the classification from the automatic classifier.

Next we will talk about various research works using rule based and machine learning approaches.

2.3.1 Rule Based Approaches

One of the first natural language processor to be integrated with clinical information systems was a rule based system called MedLee [8] [9]. It was developed in 1994 to extract clinical information and present in a structured form and was initially used to process radiology reports of chest cancer. MedLee has three stages of processing-the first one being parsing of sentences based on semantic rules and grammar, the second one is phrase standardization to reduce variation and the third one is encoding, to map the concepts to a controlled vocabulary. The system was tested on 230 radiology reports and achieved a recall and precision of 70% and 87% respectively.

BI-RAD lexicon was developed for standardizing radiology reports for breast cancer. However, most reports suffer from inconsistency and missing data. To address this issue, Nassif and his colleagues [21] developed a rule based system in 2009 to extract BI-RAD related features from mammography study. The system consists of 3 steps-syntax analyzer for processing the input into sentences, concept finder to detect BI-RAD concepts and negation detector checks for negation. The system was tested on 100 radiology reports manually tagged by radiologists and it resulted in a precision of 97.7% and recall of 95.5%.

Based on the findings from the breast imaging, a BI-RAD category is assigned to each report representing the malignancy rate of the lesion. Extracting these BI-RAD categories for structural reporting is an important task. Sippo and his colleagues [31] developed a rule based NLP system in 2013 to extract BI-RAD final assessment category from radiology reports. They tested their system on over 220 reports for each type of study-diagnostic mammography, screening mammography, ultrasound, MRI etc. achieving a recall of 100% and precision of 96.6%.

Though rule based approach achieved very good results as shown in the above work, it has a lot of limitations. It is difficult and time-intensive because it requires many experts to model the rules carefully so as to fit all the possible cases. If something goes wrong in the rules, then it does not match any of the cases and accuracy decreases by a huge margin. To fit more cases, old rules need to be extended and new ones need to be created which ends up making the rules very complex and unmanageable. Most often the rules start contradicting each other. Additionally, rules need to be changed based on the different types of hospital information systems involved which decreases its scalability and also makes it cost-inefficient.

2.3.2 Machine Learning Approaches

In this section, we will talk about use of different machine learning algorithms in the domain of radiology. The section starts with a brief overview of machine learning and its applicability in radiology. This is followed by introduction of supervised machine learning where the algorithms are trained on labeled data. Supervised learning subsection has three different paragraphs and each paragraph talks about a separate algorithm. It starts with decision tree, followed by Maximum Entropy Model and finally in the last paragraph, Support vector machine (SVM) and Conditional Random Field (CRF) are discussed. The next subsection talks about unsupervised learning, where the algorithms are trained on unlabeled data. The section ends with discussion on research work on Deep Learning.

Overview

Machine learning approach overcomes most of the limitations shown by rule based approach. Built on statistical models, machine learning gives a system the ability to learn from complex raw data and predict pattern in unknown data. In radiology, machine learning has lots of applications like early diagnosis of disease, medical image analysis, image reconstruction, language processing in reports etc.

Wang and Summer in their survey in 2012, discussed the use of machine learning in radiology and looked into the previously mentioned applications [38]. The author summarized a few advantages of using machine learning in radiology-one of them being labour saving. Due to increasing number of reports and images over the years, the workload of radiologists is increasing and becoming too much for radiologists to handle. Machine learning systems can be trained to identify complex patterns and help radiologists with the labour intensive work, so that radiologists can focus more on the high-level work. Another advantage is it was observed that many machine learning system's performance was comparable to humans and some of them were performing as good as the expert radiologists. Machine learning can be used to gain new insights into the data for example which disease gains prominence over a certain period of the year under what conditions. It is hard for humans to look into huge amount of data and answer these types of questions.

Supervised Learning

One of the very early works of using machine learning approach in the domain of radiology was performed in 1993 by Zigmond [44] who developed a software called RadTRAC to monitor follow-up of the patients from the free-text chest X-ray reports. He used dictionary based approach to identify findings related to malignancy from the reports and used machine learning approach, called *decision tree* (CART), to categorize the reports into two categories-medical follow-up required versus no medical follow-up required. The RadTrac system achieved a sensitivity of 90% and a specificity of 82% when tested on 470 radiology reports.

A more recent work done in 2013 was about extracting clinically important recommendation from radiology reports so that clinicians & other concerned persons do not miss upon any important recommendations/advices suggested for the patients by the radiologists [40]. The authors developed a recommendation extraction pipeline consisting of section segmentation, sentence segmentation and recommendation extraction. They used UMLS in feature extraction stage to match the free-text from the reports to concepts in UMLS and *Maximum entropy model*, a supervised machine learning algorithm, for feature processing. The model was tested on 800 radiology reports achieving an f-score of 0.758. This work is a continuation of another work in 2011 [41] by the same authors where they address the same aim but using rule based method to identify section boundaries whereas in this work, they used Maximum entropy algorithm to identify section boundaries. The motivation of the authors behind using machine learning for section identification was generalizing the section identification rules, which were only specific to reporting style of their institution in their 2011 work. Though the work of 2013 improved automation, the former work using rule based achieved a better f-score (87%) than the latter work.

For named entity recognition, *Conditional random field (CRF)* [18] is usually used with some variation by many researchers. Li and his colleagues [20] did a comparative study between *SVM* and CRF for disease named entity recognition and concluded that CRFs (f-score:0.86) outperformed SVMs (f-score:0.64). Torii [37] investigated the performance of CRF taggers for extracting clinical concepts and also tested the portability of this kind of tagger on different kinds of dataset. Along with CRF, the authors also used dictionary look up from UMLS for matching concepts. A master's thesis work was conducted by Joost Timmerman [36] on structuring of free-text radiology reports. He applied LC-CRF (Linear Chain CRF) for named entity recognition and achieved an f-score of 89.3%. The next work will talk about cascaded multi-stage systems, where CRF is used in multiple levels for multi-level named entity recognition. Esuli and his colleagues [7] developed a cascaded 2 stages LC-CRF, one stage CRF for identifying entities at clause level and another one at word level. They also compared it with another approach-a confidence weighted ensemble method that combines two types of classifier (standard token level LC-CRF and the cascaded 2 stage classifier mentioned in the last line) and sums up their result with equal weight. Their system was tested on 500 mammography reports and the former cascaded system performed slightly better (f-score:0.873) than the latter (f-score:0.858). Both of their systems outperformed their baseline model of standard one level LC-CRF (f-score:0.846).

Unsupervised Learning

One of the disadvantages of machine learning is requirement of labeled training corpus for supervised machine learning. In unsupervised machine learning, no manual annotation is required and the machine infers the hidden structure in data on its own. In 2013, Zhang and Elhadad did a research on biomedical named entity recognition using unsupervised approach which does not require annotated data, rules or heuristics [43]. Their system performs entity detection using a noun phrase chunker followed by a filter based on inverse document frequency and entity classification is done using distributional semantics. They tested their system on i2b2 and GENIA corpora and found that their system outperformed a dictionary matching approach.

Deep Learning

Recently, a lot of research has been going in the field of deep learning. Researchers were applying deep learning to image analysis in the beginning which has now been extended to text. For the first time, bidirectional LSTM CRF(Bi-LSTM-CRF) was applied on text data for sequence tagging by Huang and his colleagues [14]. The bidirectional LSTM component helps in looking into the past and future features and CRF looks into the sentence level tags. Their system achieved a f-score of 84.26% on named entity recognition task tested on CoNLL2003 dataset. Another very recent work was done in 2017 by a group from Stanford university, who used deep learning convolution neural network (CNN) for classifying free-text radiology reports [4]. They applied their proposed method to extract pulmonary embolism findings from thoracic computed tomography (CT) reports and compared it with a traditional NLP model, peFinder [3]. They observed that the CNN model (f-score:0.938) outperformed the peFinder model (f-score:0.867).

2.4 Summary

As conclusion of the related work section, the following important things should be noted:

1. Radiology reports need to be concise, clear and understandable for proper communication of knowledge to the outside world and for proper diagnosis of patients.

2. Structured reporting style is preferred over free-text style by many radiologists. But structured reporting should not be impose on the radiologists. Structured reporting should be such that it does not lower the accuracy of the reports.

3. Two natural language processing approaches used for radiology reports analysis are rule based and machine learning. We use machine learning approach for our purpose because in this approach, the algorithms are trained automatically to recognize patterns unlike rule based system, where the rules are handcrafted by experts. We did not want to overburden the radiologists with the task of rule creation.

4. As seen from the literature, Conditional Random Field is the best performing algorithm for sequence labeling and therefore, we use this algorithm in our project.

5. In one very recent work, deep learning performed very well on radiology reports, but deep learning models require a lot of data to get trained. Because of availability of limited labeled data, we will not be able to use deep learning for our task.

Chapter 3

Theoretical Background

In this section, we give an overview of machine learning models used in this project. We also give an overview of the radiology reporting standard for breast cancer.

3.1 Machine Learning Overview

Machine learning is a technique used to make the systems learn from the data, using statistical techniques without explicitly creating rules. Through various machine learning algorithms, these trained systems are used to make predictions on the data. There are two machine learning approaches – supervised and unsupervised. The main difference between these two approaches is that in supervised learning, systems are trained from labeled data whereas in unsupervised, no labeled data are provided. These approaches are explained in details in the next sub-sections.

3.1.1 Supervised Learning

Supervised learning has a input token (X) and a desired output variable (Y) corresponding to it and an algorithm is used to learn a mapping function from X to Y (Y=f(X)). The output variables are also known as labels or classes. The goal is to optimize this mapping function (also known as inferred function) to create model to be applied on unseen data to predict the output variables. During training, human-labeled data of input (X)-output (Y) pairs are provided to optimize the inferred function f. Parameters of this function are learned from the already labeled input-output pairs and this function is provided to unseen data for output variable prediction. An optimal scenario is to produce a function which will correctly classify all unseen data. Supervised learning can be grouped into classification problem and regression problem. In classification problem, the output variable is a category (discrete value), e.g. spam or not spam. In regression, the output variable takes continuous value e.g. all decimal values in the range 0 to 100.

3.1.2 Unsupervised Learning

Unsupervised learning is a approach of training a system to infer the underlying structure of data without training the system on labeled data beforehand. In this approach, since there are no labeled data, there is no straight forward way of evaluating the accuracy of the model. Unsupervised learning can be grouped into clustering, anomaly detection, association mining and latent variable models. In clustering, the dataset is grouped according to similarity such that objects belonging to one group are more similar to each other than objects in other group. In anomaly detection, unusual data points in the dataset are automatically discovered. This can be used in identifying fraudulent transactions. Association mining is used to discover rules that describe portions of the data. These rules can help in identifying items that frequently occur in the dataset e.g. X and Y are often bought together in supermarket. Latent variable models decomposes the dataset into multiple components or variables. It relates a set of observable variables to a set of latent variables and these observable variables are estimated using the latent variables.

3.2 Classification Task

Classification is a task of identifying to which set of categories, a new observation (input variable) belongs. For example, given a input sequence of email, x, the classification task will be to classify it into y ϵ ["spam", "not spam"]. The formula below shows the classification task mathematically:

$$y^* = \arg\max_y P(y/x)$$

The label y with the maximum posterior probability is selected as the most probable label for the input variable x. y* is the label with maximum posterior probability and the final label assigned to x.
Classification task is a type of supervised learning problem. There are many supervised learning algorithms used for this task like Naive Bayes, Support Vector Machine, Random Forest, Conditional Random Field. These algorithms will be explained in Section 3.3. Binary and multiclass classification are the two types of classification problem, which are discussed next.

3.2.1 Binary and Multiclass Classification

In binary classification, the output variable is a set of 2 classes and an instance can be classified into one of the two classes. For example, given symptoms of a patient, the patient either has a disease or not. Here, the output variable consists of two classes – has disease or no disease. Decision tree, SVM, naive bayes are some of the existing binary classifiers.

When the output variable is a set of >2 classes, it is called multiclass classification. Here, an instance can be classified into one of three or more classes. For example, given some features of a fruit, the fruit can be orange, apple, grapes or mango. Multiclass classification should not be confused with multi-label classification, where multiple classes need to be predicted for an instance. There are several methods to handle multiclass classification, two of them are i) Transformation to binary ii) Extension from binary classification algorithms.

In transformation to binary, the multiclass classification is divided into multiple binary classification problem, which is further handled using one-vs-rest or one-vs-one strategies. One-vs-rest involves training a binary classifier for each class that learns to distinguish between that class (considered as positive class) and all the other classes (considered as negative classes). At testing time, the confidence scores from all the binary classifiers are calculated and the class with the highest score is selected. One-vs-one strategy trains k(k-1)/2 binary classifiers, where k is the number of the classes. Each classifier is trained to distinguish between two classes and at the testing time, a voting scheme is applied, where the class with the highest number of votes get selected.

Another method to handle multiclass classification is extension from binary classification algorithm. Many binary classification algorithms are naturally built to handle multiclass classification like decision tree, naive bayes, neural networks. SVM is inherently a binary classification algorithm but some multiclass extensions have been developed for SVM [5].

3.2.2 Sequence Labeling

Sequence labeling, a type of structured prediction, is task of assigning categorical values to each member of a sequence of observed values. This can be seen as a classification task for each member of the sequence. Depending on the problem, the output class for each member of the sequence can be any item from a set of 2 or more classes. Thus, sequence labeling is a type of binary or multiclass classification problem. An example can be part of speech tagging, where a part of speech like noun phrase, verb phrase, is provided to each word in a sentence. In a sequence labeling task, any classification algorithm mentioned in previous section can be used. But to improve accuracy and for better prediction of labels, the labels of neighbouring members in the sequence are taken into consideration while prediction. This helps in predicting the best label sequence globally. For this purpose, some other algorithms were developed for sequence labeling, which are Hidden Markov Model, Maximum Entropy Markov Model and Conditional Random Field. These algorithms make a Markov assumption, which means the choice of label of one word is directly dependent on the labels of the adjacent words. It has been seen that Conditional Random Field perform the best in sequence labeling task, so we will use this algorithm to address one of our research questions. This algorithm is explained in detail in the next section.

3.3 Machine Learning Algorithms for Classification

In this section, the supervised learning algorithms used in our project – Naive Bayes, Support Vector Machine, Random Forest and Conditional Random Field are explained in detail.

3.3.1 Naive Bayes

Naive Bayes (NB) is a supervised learning algorithm based on Bayes theorem with naive independence assumption between the features. The algorithm is modeled such that it assigns class labels to problem instances by calculating the probability of the class label given features of the problem instance. The naive independence assumption is that the value of a feature is independent of the other features given a class variable. For example, if a house is characterized by features like has windows, a door and rooms, then Naive Bayes classifier assumes that the features – window, door and room contribute independently to probability of something being a house, irrespective of any correlation among the features. The Naive Bayes classifier is modeled as a conditional probability problem,

$$p(C_k|x_1,\ldots,x_n)$$

which means, the probability of a output class C_k given feature vector, x={$x_1,\ldots,x_n$}. C is a set of output with k possible classes, C={$C_1,\ldots,C_k$}. Using Bayes theorem, the conditional probability is decomposed as

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{3.1}$$

The terms in above equation are named as follows:

$$posterior = \frac{prior \times likelihood}{evidence} \tag{3.2}$$

The numerator of equation 3.1 is basically joint probability $p(C_k, x_1, \ldots, x_n)$.

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\ &= p(x_1|x_2, \ldots, x_n, C_k)p(x_2|x_3, \ldots, x_n, C_k) \ldots p(x_{n-1}|x_n, C_k)p(x_n|C_k)p(C_k) \end{aligned} \tag{3.3}$$

Using Naive Bayes independence assumption,

$$p(x_i|x_{i+1}, \ldots, x_n, C_k) = p(x_i|C_k) \tag{3.4}$$

which makes equation 3.3 as,

$$p(C_k, x_1, \ldots, x_n) = p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{3.5}$$

Thus, the conditional probability of Naive Bayes classifier can be written as:

$$\begin{aligned} p(C_k|x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &= \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \end{aligned} \tag{3.6}$$

where Z is the evidence, $p(x) = \sum_k p(C_k)p(x|C_k)$.
After calculating the conditional probabilities of all the classes in C, the decision of the most probable class for the feature vector is taken according to maximum a posteriori (MAP) rule (Equation 3.7), which is the class with maximum probability.

$$\hat{y} = \arg\max_k p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{3.7}$$

The prior and the likelihood are two parameter of Naive Bayes classifier that needs to be estimated. The prior is estimated in a simple way by making the classes equiprobable or weighing it based on the number of the samples of that class. The likelihood is estimated based on distribution of features called event model of the classifier. Some of the event models for discrete features are Bernoulli Naive Bayes and Multinomial Naive bayes and for continuous feature, Gaussian Naive Bayes.

In Bernoulli, the features are represented as binary attribute. For example, for a document, the word features are represented as whether the words occur or not occur in the document (1 for occur, 0 for not occur). In Multinomial, the word features are represented by their number of occurrence in the document. For Gaussian, the features are distributed according to Gaussian distribution. The likelihood probability is calculated according to Normal distribution formula. In spite of the simplified assumption of Naive Bayes, Naive Bayes works quite well in classification task like spam filtering and document classification [42].

3.3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm for binary classification. Given a set of training examples belonging to one or the other of the two classes, SVM trains a model to label a new data into any one of these two classes. One class can be referred to as positive and the other as negative. SVM classifier constructs a hyperplane that divides the positive and the negative classes. A good separation is a hyperplane that has the largest distance from the nearest training data of either class, as large distance corresponds to less generalization error of the classifier. This creates a linear SVM classifier and the hyperplane is called maximum margin hyperplane. Figure 3.1 shows the maximum margin hyperplane separating the two classes of data.

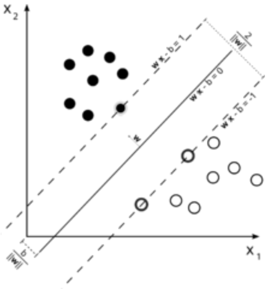


FIGURE 3.1: Support Vector Machine showing maximum margin hyperplane (taken from [34])

Suppose the data looks like this $(\vec{x_1}, y_1) \ldots (\vec{x_n}, y_n)$, where x_i is the data point belonging to class y_i (y_i can only be -1 or 1). In a linear SVM, we want to find the maximum margin hyperplane which divides the data point having class y_i=1 from the data point having class y_i=-1. A hyperplane is defined as $\vec{w}.\vec{x} - b = 0$, where w is the vector perpendicular to the hyperplane, $\vec{x} = \vec{x_1}, \ldots, \vec{x_n}$ is the feature vector and b is the position of the hyperplane.

Linear SVM can be divided into two types – hard margin and soft margin. Hard margin is used if the training data is linearly separable. Two parallel hyperplanes are selected that separate the two classes and the distance between the hyperplanes is as large as possible. In this case, the maximum margin hyperplane lies half way between these two hyperplanes. The search for maximum margin hyperplane in hard margin is modeled as the optimization problem of minimizing $\frac{1}{2}||w||^2$ subject to $y_i(\vec{w}.\vec{x_i} - b) \geq 1$. Soft margin is used when the data is not linearly separable. Soft margin introduces a hinge loss function, $max(0, 1 - y_i(\vec{w}.\vec{x_i} - b))$. If x_i lies on the correct side of the margin, then $y_i(\vec{w}.\vec{x_i} - b) \geq 1$ and the hinge loss functions results in zero. If the data x_i is on the wrong side of the margin, then the following function is minimized, $[\frac{1}{n}\sum_{i=0}^{n} max(0, 1 - y_i(\vec{w}.\vec{x_i} - b))] + \lambda||\vec{w}||^2$. Nonlinear SVM is also used for data which are not linearly separable, by applying kernel trick to maximum margin hyperplane.

Once the SVM classifier is trained using the labeled data, the score for classifying an unseen data (represented using feature vector, $\vec{x}$) is calculated using the function $f(\vec{x}) = \vec{w}.\vec{x} - b$. The result of the function decides the class of the unseen data. If

$f(\vec{x}) > 0$, then the $\vec{x}$ is labeled as positive class or else negative class. As SVM is a binary classifier inherently, for multiclass classification problem, techniques discussed in Section 3.2.1 are used.

3.3.3 Random Forest

Random Forest (RF) is an ensemble of many decision trees, where each decision tree is fitted on a random sub-sample of the training set and the model outputs the class which is the the mode of the class output by individual tree. Random forest model follows the idea of bagging or bootstrap aggregating, which means the use of multiple classifiers to make a decision and ultimately choosing the class with maximum votes. Construction of each tree in the random forest model is according to the following algorithm. Let the number of training sample be N and number of features be M. All the features are not used to construct the tree, rather, a random selection of m features is done, where $m < M$. Only these m features are used to calculate the best split at each node of the decision tree. A training set is built by choosing N times with replacement from all N available training samples. This is called bootstrap sample. Also, not all the bootstrap samples are used to build the tree. About one-third of the bootstrap samples is left out to estimate the classification error rate of the tree on the predicted classes and also to measure the importance of the feature variable. This left out data is called out-of-bag (OOB) data. Unlike the normal decision trees, each tree is fully grown and not pruned. After all the trees are build and the forest is completed, a new sample can be classified by taking the majority vote among all the tress in the forest (resembling the bootstrap aggregating idea). Random forest learning is quite fast and it can handle a large amount of input variables.

3.3.4 Conditional Random Field

Sequence labeling, as discussed before, is a type of problem where the sequence of classes y=$(y_1, \ldots, y_n)$ has to be predicted given a sequence of observed feature vectors, x=$(x_1, \ldots, x_n)$. Graphical models are used to model the dependencies between the observed variables. A type of graphical model, called generative model, uses joint probability distribution p(y,x) over x and y to model the problem. Ab example of generative model is Hidden Markov Model. But, this approach has some limitations. If we want to model using joint probability, then we have to model the input features p(x), which has complex dependencies among them. Modeling the dependencies among the features can make the problem intractable and ignoring the dependencies results in reduced performance.

A solution to this problem is directly modeling the conditional probability p(y | x), which is the only thing needed for classification. This is the approach taken by Conditional Random Field (CRF) [18]. Conditional Random Field has the advantage that the dependencies that occur in x play no role in the conditional model, so conditional model has a much simpler structure than joint model. CRF was developed by John Lafferty, Andrew McCallum and Fernando Pereira in the year 2001, as a framework to build probabilistic models to segment and label sequence data, and a better alternative to the Hidden Markov Model (HMM) and Maximum Entropy Markov Model (MEMM). CRFs take into account the context of the token, which means the classes of neighbouring tokens. Ordinary classifiers like Naive Bayes base their prediction only looking at features of single instances and not its surrounding labels. Thus, CRFs are preferred for structured predictions. CRFs provide several advantages over HMM including the ability to relax strong independence assumptions made in those models. CRFs also avoid the fundamental limitation of MEMM, that is bias towards states with few successor states, known as label bias problem. Figure 3.2 shows how CRF is related to HMM, Naive Bayes and Logistic Regression.

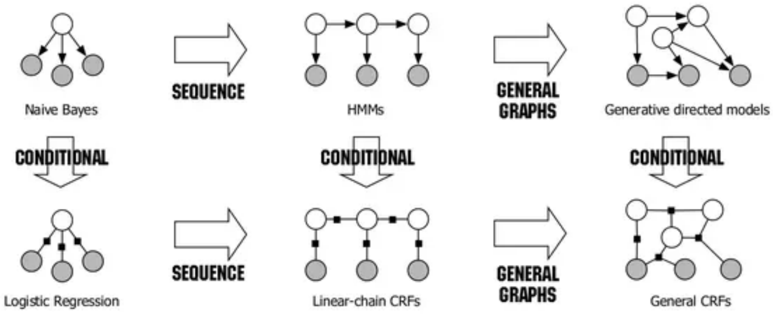


FIGURE 3.2: Diagram of relationship among Naive Bayes, Logistic Regression, HMM, LC-CRF and general CRF (taken from [35])

In the original paper [18], CRFs are defined as follows:
Let G=(V,E), E are edges, which are the cliques of the graph and V are the vertices $Y = (Y_v)_{v \epsilon V}$, such that Y is indexed by vertices of the graph G.
Then, (X,Y) is a conditional random field in case, when conditioned on X, the random variable Y_v obey the Markov property with respect to the graph, $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means w and v are neighbours in the graph.

Thus, CRFs are undirected graphs that can be divided into X and Y disjoint set, where X is the observed input features and Y is the output variable, from which conditional probability p(X | Y) can be modeled. Since, conditional probability is used to model CRF, CRFs belong to the class of discriminative models. A type of CRF is called Linear Chain CRF, which will be described next.

Linear Chain Conditional Random Field

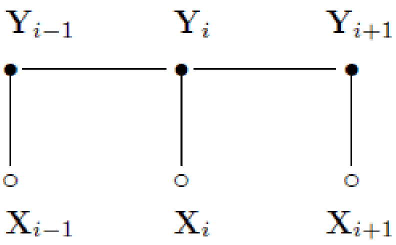


FIGURE 3.3: Linear Chain Conditional Random Field (taken from [18])

Though, CRFs have arbitrary structure (general CRFs shown in Figure 3.2), there is a type of CRF which supports sequence modeling, called Linear Chain CRF (LC-CRF) (shown in Figure 3.3). These are used in text processing, bioinformatics and computer vision. Given a sequence of text, $X = (x_1, \ldots, x_n)$, where x_i is a token feature vector and $Y = (y_1, \ldots, y_n)$ are the corresponding labels or classes, where y_i is the label for feature vector x_i. LC-CRFs model the conditional probability p(Y | X) according to the following formula:

$$p(Y|X) = \frac{1}{Z(X)} \prod_{i=1}^{|X|} \Psi_u(y_i, x_i) \prod_{i=1}^{|X|-1} \Psi_b(y_{i-1}, y_i, x_i) \tag{3.8}$$

where i is token number in the sequence, $\frac{1}{Z(X)}$ is a normalizing factor to ensure that the resultant value is a probability, Ψ_u are factors of feature vectors x_i and label y_i, and Ψ_b are factors of feature vector x_i and labels y_{i-1} and y_i.

$$Z(X) = \sum_Y \prod_{i=1}^{|X|} \Psi_u(y_i, x_i) \prod_{i=1}^{|X|-1} \Psi_b(y_{i-1}, y_i, x_i) \tag{3.9}$$

$$\Psi_u(y_i, x_i) = \exp \sum_k \theta_k f_k(y_i, x_i) \tag{3.10}$$

$$\Psi_b(y_{i-1}, y_i, x_i) = \exp \sum_h \theta_h f_h(y_{i-1}, y_i, x_i) \tag{3.11}$$

In equations 3.10, 3.11, f_k and f_h are feature functions, f_k depends on the observation and label at current time step i and f_h depends on the observation and label at current time step i and also on label at previous time step i-1. θ_k and θ_h are the weights of the feature functions.

Training and Testing

For training, θ_k and θ_h are the parameters to be estimated from the training data and the conditional probability p(Y | X) needs to be maximized. In our project, we used a quasi-Newton gradient descent method called the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) training algorithm for parameter estimation. As, this algorithm is not the scope of study for this project, we will just give an outline of the algorithm. L-BFGS uses an estimation to the inverse Hessian matrix to search through variable space. To limit the use of computer memory, only the vectors which represent the approximation of the matrix at a time are stored instead of the n x n matrix, where n is the number of variables in the problem. The gradient of the entire training set is computed using current weights and then batch update is done to move in small steps towards the minimum along the computed gradient.

In testing, for a input sequence X, the most probable label sequence y* needs to be found by taking that label which has the maximum conditional probability p(Y | X). Generally, Viterbi algorithm is used to find the best label path in the model for a sequence of input.

3.4 Radiology Reports on Breast Cancer

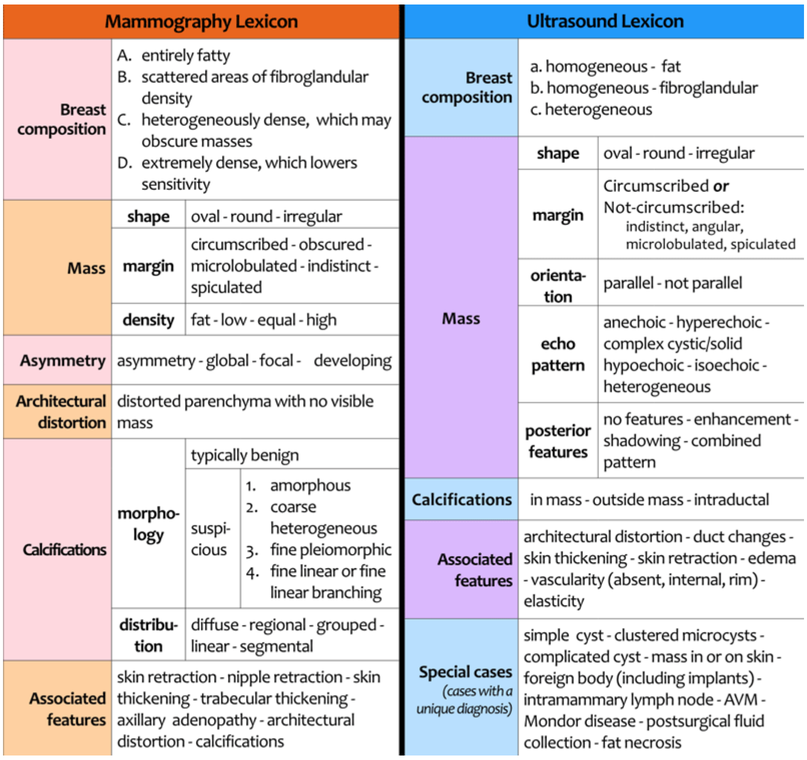
This section introduces to the medical domain of our project. We will provide with a brief overview of the structure and elements present in a radiology report on breast cancer, in accordance with ACR BI-RADS [30]. This section provides a basis for the labeling scheme followed for our project. A person going through the diagnostic process for breast cancer, may either have a benign lesion or a cancer. Radiology reports and a further check up from the clinician informs the person about his condition, which makes radiology reports quite important.

The following elements should be present in a standard reporting process:

1. Indication for examination
   This part contains the screening, diagnostic or follow-up of patient about either a benign lesion or a cancer. This part should also contain the patient's medical history.

2. Succinct description of the overall breast composition
   In BIRAD edition 2003, breast composition category was decided on the overall breast density, resulting in 4 ACR categories – ACR category 1 (<25% fibroglandular tissue), category 2 (25-50% fibrogandular tissue), category 3 (50-75%) and category 4 (>75%). In BIRAD edition 2013, the breast composition categories were changed to a, b, c and d.
   a - The breast are almost entirely fatty.
   b - There are scattered areas of fibroglandular density.
   c - The breasts are heterogeneously dense, which may obscure small masses.
   d - The breasts are extremely dense, lowering the sensitivity of mammography.

3. Clear description of important findings
   This part of the report describes the important observations noted by the radiologist from the images. Findings can occur from mammography study, ultrasound or Magnetic Resonance Imaging (MRI). Figure 3.4 gives an overview of the mammography and ultrasound lexicon. We will discuss mammography study in details because our project focuses on reports from mammography. Note: The images produced by mammography study are called mammograms. An abnormality in the breast can be referred to as a lesion.



Mammography Lexicon			
Breast composition	A. entirely fatty B. scattered areas of fibroglandular density C. heterogeneously dense, which may obscure masses D. extremely dense, which lowers sensitivity		
Mass	shape	oval - round - irregular	
	margin	circumscribed - obscured - microlobulated - indistinct - spiculated	
	density	fat - low - equal - high	
Asymmetry	asymmetry - global - focal - developing		
Architectural distortion	distorted parenchyma with no visible mass		
Calcifications	morphology	typically benign	
		suspicious	1. amorphous 2. coarse heterogeneous 3. fine pleiomorphic 4. fine linear or fine linear branching
	distribution	diffuse - regional - grouped - linear - segmental	
Associated features	skin retraction - nipple retraction - skin thickening - trabecular thickening - axillary adenopathy - architectural distortion - calcifications		

Ultrasound Lexicon		
Breast composition	a. homogeneous - fat b. homogeneous - fibroglandular c. heterogeneous	
Mass	shape	oval - round - irregular
	margin	Circumscribed **or** Not-circumscribed: indistinct, angular, microlobulated, spiculated
	orientation	parallel - not parallel
	echo pattern	anechoic - hyperechoic - complex cystic/solid hypoechoic - isoechoic - heterogeneous
	posterior features	no features - enhancement - shadowing - combined pattern
Calcifications	in mass - outside mass - intraductal	
Associated features	architectural distortion - duct changes - skin thickening - skin retraction - edema - vascularity (absent, internal, rim) - elasticity	
Special cases (cases with a unique diagnosis)	simple cyst - clustered microcysts - complicated cyst - mass in or on skin - foreign body (including implants) - intramammary lymph node - AVM - Mondor disease - postsurgical fluid collection - fat necrosis	

FIGURE 3.4: Mammography and ultrasound lexicon (taken from [45])

Mammography findings can be divided into the following 5 parts:

- Mass: It is a space occupying 3D lesion. It can have shape, margin, density.
- Calcification: These are calcium deposit within breast tissue, appearing as white spots or flecks on a mammogram. This is described by morphology and distribution.
- Architectural Distortion: When the normal mass is distorted with no definite mass visible, then it called architectural distortion.
- Asymmetry: These are the findings that represent unilateral deposits of fibroglandulair tissue not conforming to the definition of a mass.
- Associated features: These are seen with suspicious findings like masses, asymmetry and calcification. They can be skin retraction, nipple retraction, skin thickening etc.

All the types of important findings above has an associated location, describing the area of occurrence in the breast. A location can be left, right or both breasts. Within a breast, the position of the lesion can be described according to clockface or quadrant notation e.g. upper outer quadrant, upper inner quadrant. The quadrant position are described in the Figure 3.5.

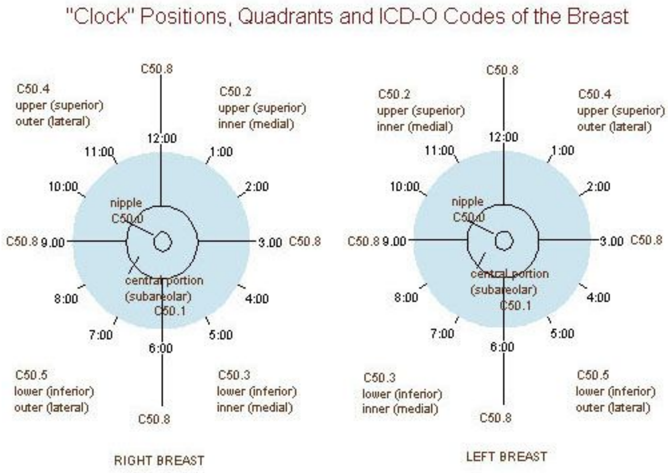
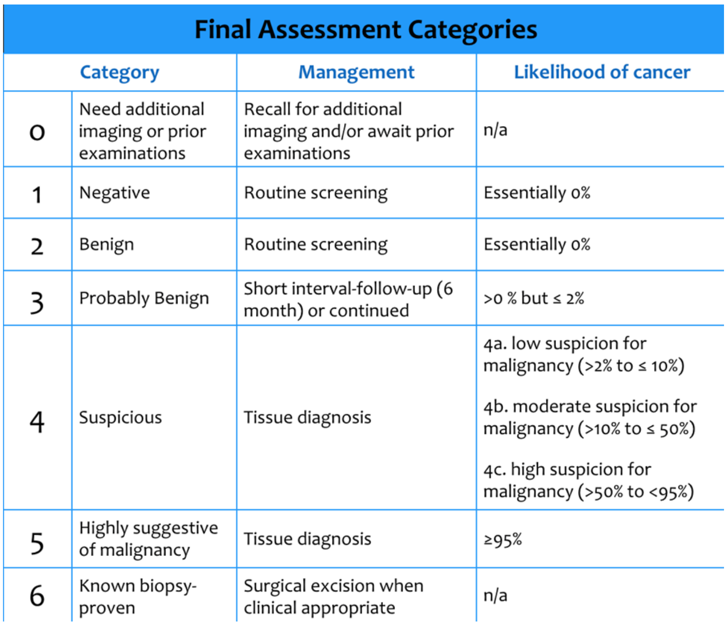


FIGURE 3.5: Quadrant position of the breast (taken from [29])

4. Comparison to previous examinations if deemed important by the interpreting physician
   Comparison with previous examinations of the patient is assumed to be important to find out if the finding is stable or is changed. Comparison is not important when the finding is inherently highly suspicious.

5. Assessment
   A final assessment category known as BI-RADS category is necessary to categorize the findings according to malignancy rate. The BI-RADS category can take any one of the values – 0, 1, 2, 3, 4, 5, 6, with 0 being benign to 6 being most malignant. Figure 3.6 gives an overview of all the BI-RADS category and their meaning.



Final Assessment Categories

Category		Management	Likelihood of cancer
0	Need additional imaging or prior examinations	Recall for additional imaging and/or await prior examinations	n/a
1	Negative	Routine screening	Essentially 0%
2	Benign	Routine screening	Essentially 0%
3	Probably Benign	Short interval-follow-up (6 month) or continued	>0 % but ≤ 2%
4	Suspicious	Tissue diagnosis	4a. low suspicion for malignancy (>2% to ≤ 10%) 4b. moderate suspicion for malignancy (>10% to ≤ 50%) 4c. high suspicion for malignancy (>50% to <95%)
5	Highly suggestive of malignancy	Tissue diagnosis	≥95%
6	Known biopsy-proven	Surgical excision when clinical appropriate	n/a

FIGURE 3.6: BI-RADS assessment category and their description (taken from [45])

6. Management
   This part contains the some management recommendations as a next step to be taken. If suspicious abnormality is found, then the report should indicate "biopsy should be performed in the absence of clinical contraindication".

3.5 Summary

Naive Bayes, Support Vector Machine, Random Forest and Conditional Random Field are some of the supervised machine learning algorithms used for any classification task. LC-CRFs are very useful for sequence labeling, which is a type of classification task. Therefore, we will use these algorithms for our classification task and sequence labeling. Radiology reports on breast cancer are structured according to ACR BI-RADS guidelines and we will follow these guidelines to develop a labeling scheme for our radiology reports.

Chapter 4

Methodology

The approach section in the research paper in part one already contains the methodology in detail. This section contains some extra information about the dataset, the features used in heading and content identification task and some information about the hierarchical Model B that is missing from the paper.

4.1 Dataset

5000 free-text radiology reports on breast cancer ranging over the years 2012 to 2017 were collected from the radiologists and innovation manager working at ZGT hospital, Netherlands. The dataset collected consists of the following information:

TABLE 4.1: Attributes in the dataset and their description

Attributes	Description	Attributes	Description
Geslacht	Gender of the patient	Gebdat	Date of birth of the patient
Onderzdat	Examination Date	Verslag	Radiology report containing clinical data, findings, conclusion
Conclusie	Conclusion of the report – also contains the BIRAD score	Omschr	Description of procedure
Rontverrid	Reference number to image	Onderznr	Examination number
Code	Code of procedure	Omschr-kamer	Room of examination
Indicatie	Clinical Data	Specialism	Type of physician examining the patient

For our project, we will only work with the verslag (report) attribute which contains the radiology report. The radiology report also contains the indicatie and conclusie attributes within it.

4.2 Heading and Content Identification

Heading identification is a multiclass classification problem, where the sentences are to be classified into any one of the following classes: *heading*, *not heading* and *title*.

Content identification is a multiclass classification problem, where the sentences are to be classified into *title*, *clinical data*, *findings* and *conclusion*. Sometimes the report contains a sentence mentioning the name of the radiologist who wrote the report. Those sentences were labeled as *names* as they could not be put into any of the above mentioned sections. For our task, the *names* class is not that important.

4.2.1 Manual Annotation

Out of 5000 reports, 180 reports were manually labeled for building the machine learning models. The reports were separated into sentences by splitting at newline ('$\backslash n$') and manually annotated into three classes – title, heading and not heading for heading identification and into five classes – title, clinical data, findings, conclusion and names for content identification. Table 4.2 shows a sample report split into sentences in column 1, column 2 shows the labels for heading identification (HI) and column 3 shows labels for content identification (CI).

TABLE 4.2: Example of manual annotation of a radiology report for heading and content identification

Sentences	**Labels(HI)**	**Labels(CI)**
Verslag - Mammografie follow up bdz - 15-11-2016 09:50:00:	Title	Title
Klinische gegevens:	Heading	Clinical Data
Screening ivm familiaire belasting mammacarcinoom,	Not Heading	Clinical Data
Verslag:	Heading	Findings
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,	Not Heading	Findings
Conclusie:	Heading	Conclusion
BIRADS-classificatie twee, Stationair beeld, Geen maligniteitskenmerken,	Not Heading	Conclusion

4.2.2 Features Used

Term Frequency Text Representation

After preprocessing, the words generated from the reports were used as features for the classifiers. A vocabulary was built for all the unique words in the radiology reports, after preprocessing, and each of the words was represented by a unique number. A document-term matrix was created for all documents (sentences in our case). A document-term matrix has the sentences as rows and all the unique words in the vocabulary as columns. Each sentence is represented by how many times each word in the vocabulary occurs in that title. Thus, the features for the classifier are the distinct words, $[w_1, w_2, .., w_i, ..w_n]$, with the frequency of occurrence of the word w_i, in the sentence as its value.

TF-IDF Text Representation

Finding out the important words that represent a given category can be done in many ways. The straight forward solution can be to count the number of terms occurring in a given category and deciding the most frequent ones to be the words that represent the category. This is the method that was followed in the previous part (Term frequency text representation). However, this method doesn't consider the other categories. Here, TF-IDF is used which considers not only the term frequency in the considered category but also in how many documents does the term occur. TF-IDF is defined as

$$tfidf(t_k, c_i) = tf(t_k, c_i) * log\left(\frac{C}{df_i}\right) \tag{4.1}$$

where

- $tf(t_k, c_i)$ denotes number of times term t_k occurs in category c_k
- df_i denotes the number of categories containing term i
- C denotes the total number of categories.

The TF-IDF basically is based a heuristic intuition that a word/term which occurs in almost all the categories is not a good discriminator for any category, and should be given less weight than one which occurs in few categories.
Thus, TF-IDF score was used instead of term frequency to represent each word feature. TF-IDF score was calculated for all the words that was generated after preprocessing. Each word in the sentence was represented by their TF-IDF score, which was used as feature to train the model. The words not present in the sentence was given a value of zero.

Length of the Sentence

Another feature used for training the classifier is the base 10 logarithm of length of the sentence. By length of the sentence, it means the number of words in the sentence. It was observed that the headings are generally short sentences containing one or two words whereas the sentences which are not heading are longer containing a lot of words. Thus, the length of the sentences may help in categorizing the sentences.

Symbol at the End of the Sentence

The last symbol of each line was another feature for our classifier. The headings end with a colon (:) usually and the rest of the sentences either end with a comma (,) or just a letter. If the last symbol was colon, we denoted it by 1, otherwise 0.

Note: For the task of heading identification, all the above features were used. For the task of content identification, only the TF-IDF text representation feature and length feature were used. End of sentence symbol feature was not used, as sentences in two different sections usually ends with similar symbol (','), thus, not contributing any unique feature to content identification problem.

4.2.3 Classifiers

Three different classifiers were trained on our data and their performance was compared.

Naive Bayes (NB)

We used Multinomial Naive Bayes classification algorithm to built our model, as multinomial NB is useful for classification having discrete features like word frequency/ TF-IDF in text classification. It uses Bayes theorem to calculate the probability of a class given the features of the data, with the assumption that the features are conditionally independent. The training set was used to calculate the parameters like likelihood probabilities and prior for our Naive Bayes model. Classification of the sentences in the test set was done by using these parameters to calculate the posterior probability of each class given the sentences in the test set. Further, the class with the highest posterior probability was selected as the class for the sentence. Laplace smoothing was used while calculating likelihood probability to handle unseen features in the test set. MultinomialNB, a sklearn python package was used for building the model.

Support Vector Machine (SVM)

We used SVM classification algorithm to built another model. As SVM works by finding a separation between hyperplanes of different classes of data, its learning ability is independent of the dimensionality of the feature space. Thus, SVM works well for text classification, which has high dimension feature space from huge vocabulary size. For implementation of SVM, we used SGDclassifier from sklearn python package. Linear SVM classifier with soft margin was used. Stochastic gradient descent was used for updating the loss function and alpha, the constant that gets multiplied with regularization term was set as 0.001. Learning rate was set as optimal, which uses the value of alpha to fix its value. To prevent overfitting, L2 regularizer was used to shrink the model parameters towards zero vector. The seed of random number generator for shuffling the data was set as 42 and maximum iteration over the training data was set as 5. The rest of the parameters were set to default value.

Random Forest (RF)

A third model was built using Random Forest classifier. Random Forest is an ensemble of decision trees, where each decision tree is fitted on a random sub-sample of the training set. After the training phase is complete, a new data is classified by taking the majority vote amongst all the trees. For implementation of Random Forest, we used RandomForestClassifier from sklearn python package. The maximum depth of the true was set as 10 and random state was set as 0. Bootstrap was set as True to select sub-samples with replacement. All other parameters were set to their default value.

4.2.4 Training and Testing

5 fold cross validation technique was used for training and testing the classifiers on 180 manually labeled reports. This technique was adopted because there were not many manually labeled reports available, thus, the most efficient thing was to use all

the available reports both for training and testing. The 5 fold cross validation technique divides the labeled dataset into 5 parts and uses 4 parts to train the classifier and the remaining 1 part to test it. It repeats the same thing 5 times, each time taking a different set of 4 parts of the dataset for training and the other 1 part for testing. This helps in using the whole set of labeled data for training and testing but never training and testing on the same part of data in the same iteration.

4.3 Automatic Structuring

Automatic structuring of free-text mammography findings is a sequence labeling problem. The problem definition and all the other details are described in the paper. Here, we explain a bit more about the manual annotation and the hierarchical model B.

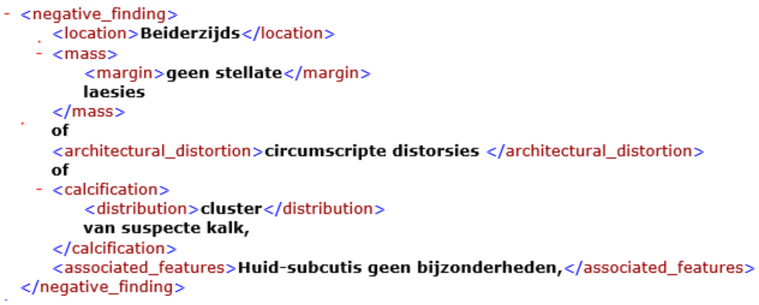
4.3.1 Manual Annotation

TABLE 4.3: First, second and third level classes for automatic structuring

First Level	Second Level	Third Level
breast composition	-	-
positive finding	mass	location margin density shape size associated features
	calcification	location morphology distribution size associated features
	asymmetry	location size associated features
	architectural distortion	location associated features
	associated features	location
negative finding	location	
	mass	location margin
	calcification	location morphology distribution
	asymmetry	location
	architectural distortion	location
	associated features	location

Based on the ACR BI-RADS guidelines, a 3-level annotation scheme was developed along with two radiologists. The report was first annotated into positive finding, negative finding and breast composition, then the positive finding and negative finding were annotate into mass, calcification, asymmetry, architectural distortion and associated features. At the third level, these findings were annotated into classes important for each of these findings. The complete class of a token is referred as the global class, which is constructed by concatenating classes from each level. It is represented as Class_Level1/Class_Level2/Class_Level3, where Class_Level1 is the class of the token from level 1 and so on. Some tokens may not have a prediction at the second or the third level, in which case the global class looks like Class_Level1 or Class_Level1/Class_Level2. Table 4.3 shows classes which were used for annotation at each level.

The first level and all the sub levels of each class have a other class, which takes the tokens that do not belong to the rest of the classes at that level. For example, the second level of negative finding has a other class along with mass, calcification, asymmetry, architectural distortion and associated features. Figure 4.1 shows manual annotation of the negative finding of a report. Not important words like 'of' ('or' in English) connecting classes like mass and architectural distortion are labeled as other within negative finding. Note that, 'of' is not enclosed within a 'O' tag (other) inside negative finding. Our annotation was such that if a token is not within a tag at the second and third level, then it is assumed to be a other tag. But tokens at the first level are explicitly annotated as 'O'. At third level, positive finding has associated features class under mass, calcification, asymmetry and architectural distortion. There is also another associated features class at the second level of positive finding. The associated features class at the second level is a separate finding in itself and the associated feature at third level is observed in association with another finding like mass, calcification, asymmetry and architectural distortion.



```
- <negative_finding>
    <location>Beiderzijds</location>
  - <mass>
        <margin>geen stellate</margin>
        laesies
    </mass>
    of
    <architectural_distortion>circumscripte distorsies </architectural_distortion>
    of
  - <calcification>
        <distribution>cluster</distribution>
        van suspecte kalk,
    </calcification>
    <associated_features>Huid-subcutis geen bijzonderheden,</associated_features>
  </negative_finding>
```

FIGURE 4.1: Example of annotated negative finding of a report

On inspection, some of the things found about the dataset are as follows:

1. All the classes at the 3 levels do not occur in every report. Generally, other and breast composition occur in almost every report.

2. The order of appearance of the classes in the reports are not fixed, but breast composition usually occur at the beginning, followed by positive finding and then by negative finding. Many times, the position between positive finding and negative finding get interchanged.

3. There may be multiple occurrence of a particular class in a report.

4. One token can have multiple tags at the same level, for example, the token 'kalk' can be associated with negative finding and also with positive finding.

4.3.2 Hierarchical CRF with Combined Classes

This model, referred to as Model B, is already explained in the paper. Here, some more details about the model will be given which could not explained in the paper. The aggregated classifiers are given only that piece of text, which may contain the class to be predicted by the aggregated classifier. It may not be that apparent which higher level classes are passed to the aggregated classifiers in our model and the results may vary if it is passed differently. This is why it is required to explicitly mention it. The following part will explain which higher level classes are passed to the aggregated classifiers:

- CB-5 (location) takes as input, tokens classified as positive finding and negative finding. This is because location can be found as a third level class within all the second level classes of positive finding and negative finding. It can also be found as a second level class with negative finding. So, basically, anything classified as positive or negative finding, can have a location.
  CB-5 could have also taken input differently, for example, tokens classified as mass, calcification, architectural distortion, asymmetry and associated features can be given as input. This would lead to different surrounding tokens and thus a bit different prediction.

- CB-6 (margin) takes as input, tokens classified as mass, as both mass of positive finding and negative finding have the class margin at the third level.

- CB-7 (morphology and distribution) takes as input, tokens classified as calcification, as both calcification of positive finding and negative finding have the class morphology and distribution at the third level.

- CB-8 (associated features) takes as input, tokens classified as positive finding at the first level and mass, calcification, architectural distortion and asymmetry at the second level. This classifier is not used to predict the associated feature under negative finding class, as the associated feature under negative finding is written very similarly in every report, leading to already very good prediction. CB-8 is also not used to predict associated feature at the second level under positive finding (PF/AF). The reason is we wanted to use PF/AF's context for its prediction. Also, we created the aggregated classifiers mainly to predict the third level classes and as PF/AF is a second level class, we did not used CB-8 for its prediction.

- CB-9 (size) takes as input, tokens classified as positive finding at the first level and mass, calcification and asymmetry at the second level.

Chapter 5

Experiments and Discussion

In this section, details about the experimental setup are discussed at first. This is followed by results from heading and content identification and automatic structuring of mammography findings. The paper in part 1 of this thesis already contains experimental setup and some important results. These section only contains the results not given in the paper and also extra information on experimental setup.

5.1 Experimental Setup

Evaluation metrics are required to evaluate how a classifier is performing. Confusion matrix is generated for the each of the classifiers and the true positive (TP), false positive (FP), false negative (FN) and true negative (TN) values for each class are calculated. These values are used to calculate evaluation metrics, precision, recall and F_1 score. Precision (p) is a measure of the number of actually correct instances among the instances identified as correct by the system. Recall (r) is a measure of the number of actually correct instances classified by the system among all the correct instances. F_1 score of a class is the harmonic mean of precision and recall of that class.

$$p = \frac{TP}{TP + FP}$$

$$r = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2pr}{p + r} = \frac{2TP}{2TP + FP + FN}$$

TP, TN, FN and FP are binary classification concepts. As heading and content identification and automatic structuring are multiclass classification problem, these measures were calculated by following one-vs-rest binary classification, where the class in consideration is positive and the rest of the classes are negative. The performance of the classifier as a whole was calculated using microaveraged and weighted macroaveraged F_1 score.

Microaveraged F_1 score (F_1^{μ}): It is calculated by summing over TP of all the classes and dividing by the sum of TP, FN and FP of all classes. In the formula below, TP_i is TP of a class i and n is the number of classes.

$$\frac{2 * \sum_{i=1}^{n} TP_i}{2 * \sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i + \sum_{i=1}^{n} FN_i} \tag{5.1}$$

Weighted Macroaveraged F_1 score (F_1^{M}): It is the weighted average of the F_1 scores of all the classes. In the formula below, w_i and F_{1i} are the weight and the F_1 score of class i respectively.

$$\sum_{i=1}^{n} w_i * F_{1i} \tag{5.2}$$

5 fold cross validation was used for evaluating the models of heading identification and 4 fold cross validation was used for automatic structuring. The F_1^M and F_1^μ of each fold were weighted averaged to get the performance of the models over all the folds. The F_1 scores reported in the rest of this section are the scores of cross validation.

Confusion matrix was generated by adding the TP, TN, FN and FP values of each fold to get an overall confusion matrix for k fold cross validation. In this experiments section, two types of confusion matrix were generated – confusion matrix without normalization and confusion matrix without normalization. The former confusion matrix is the normal one with the exact number of correctly and incorrectly predicted instances for all the classes. The latter confusion matrix is another version of the former confusion matrix, where all the values in each row are divided by the sum of all the values in that row (class support size). Henceforth, the former type will be referred as confusion matrix and the latter will be referred as normalized confusion matrix. Normalized confusion matrix is used when the exact TP, TN, FN and FP values are not required, rather, a more visual interpretation of the misclassified classes is needed. For heading and content identification results (Section 5.2), confusion matrix without normalization was used and for automatic structuring (Section 5.4), normalized confusion matrix was used. The reason for the use of normalized version for the latter case is that automatic structuring has 34 labels making the visualization of misclassified classes harder in the normal confusion matrix.

For automatic structuring, we compare performance of the classes at each level for our baseline model, Model A and Model B. To elaborate, a global class of a token is designed as Class_Level1/Class_Level2/Class_Level3 as described in Chapter 4. We compared performance of classes at level 1 (Class_Level1), at level 2 (Class_Level1/Class_Level2) and global class. Model A and B have different classifiers at different levels like CA-1 at level 1 for Model A, CB-1 at level 1 for Model B, unlike, the baseline model, which has only one classifier to classify the global class of a token. Thus, to establish a comparison of Model A and B with the baseline model for the performance of classes at first and second level, we separated the predicted global class of baseline model into 3 parts – the first part as the predicted class of first level (Class_Level1), the first and second part combined as the predicted class of second level (Class_Level1/Class_Level2) and the global class. These comparisons are shown in Section 5.4.

5.2 Heading Identification

We applied different combinations of the features to train our 3 models as shown in Table 5.1. It can be seen that RF performs the best when word list and end of sentence (EOS) symbol are used as features. SVM performs best with every combination of features and NB performs best with only word list feature. TF-IDF representation of word list was used as feature. All scores are shown in F_1^M.

TABLE 5.1: Performance of classifiers in terms of F_1^M scores for different feature combinations for heading identification

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

Table 5.2 shows the performance of the 3 models for the task of heading identification using only word list feature. Heading and not heading classes were best predicted by SVM and NB with an F_1 score of 0.96 and 0.98 respectively. Title class was best predicted by RF with an F_1 score of 0.99. The performance of the model as a whole is shown with F_1^M and F_1^μ.

TABLE 5.2: Heading identification performance in terms of F_1 scores

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 ($\mathbf{F}_1^M$)	0.97	0.97	0.92	1591
Avg ($\mathbf{F}_1^\mu$)	0.97	0.97	0.92	1591

Figure 5.1 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, and most of the instances of all the classes were correctly classified.

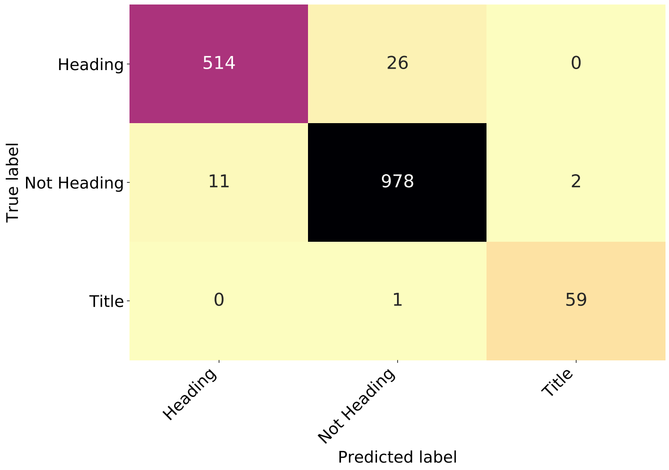


FIGURE 5.1: Confusion matrix heat map: Heading identification using SVM

5.3 Content Identification

Table 5.3 shows the performance of the 3 models on the task of content identification using TF-IDF representation of the word list as the only feature. The classes conclusion, clinical data, title and findings were predicted best by SVM with an F_1 score of 0.92, 0.94, 0.99 and 0.94 respectively. Names class which is not that important for our task, was predicted with an F_1 score of 1.0 by SVM.

TABLE 5.3: Content identification performance in terms of F_1 scores

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
Names	0.79	1.00	0.00	35
Avg ($\mathbf{F}_1^M$)	0.87	0.94	0.80	1591
Avg ($\mathbf{F}_1^\mu$)	0.88	0.94	0.81	1591

Figure 5.2 shows the heat map representation of the confusion matrix for content identification using SVM. Both the conclusion and clinical data classes were wrongly predicted as findings in 44 and 25 out of their total instances respectively. This is justified because conclusion, clinical data and findings describe a lot of things and many of the words in these three may overlap, leading to the misclassification.

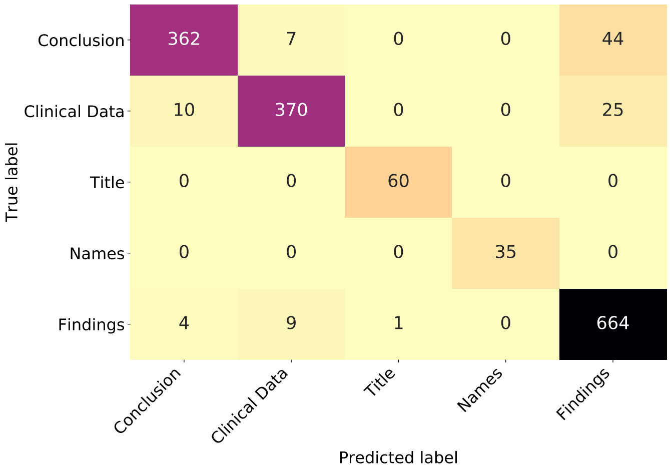


FIGURE 5.2: Confusion matrix heat map: Content identification using SVM

For the content identification task, we experimented with 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. Table 5.4 shows the performance in terms of F_1^M score for different combination of these features. SVM shows the best performance (F_1=0.94) for TF-IDF word features. This is same as the average F_1^M for SVM shown in the Table 5.3.

TABLE 5.4: Performance of NB and SVM classifiers in terms of F_1^M scores for different feature combinations for content identification

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

It can be observed from the table that NB performs much better for term frequency word list than for TF-IDF, whereas, SVM performs slightly better for TF-IDF than term frequency. Another observation is SVM performs much worse than NB when using the length feature together with word list. To explain this observation, we have to first understand that, in very simple terms, a classifier performs an optimization problem on the function, $y = f(x) = w_1.x_1 + w_2.x_2 + \ldots + w_n.x_n$, for finding the best possible output y based on the features $x_1, x_2, \ldots, x_n$. Weights such as $w_1, w_2, \ldots, w_n$ are assigned to different features such that the function gets optimized. In NB, the feature, x_i, is a conditional probability of x_i given a class, whereas, in SVM, x_i takes the actual value of the feature.

So, for SVM, if some features lie in the range of 0-1, then they have less impact on f(x) than features which lie in the range of 100. For this reason, when using combination of the length of the sentence (values in the range of 100), and the TF-IDF (values in the range of 0-1) in SVM, the length factor dominates and results in low F_1 score of 0.29. On replacing length by $\log_{10}$ of length, all the features values lie in the same range and the F_1 score improves to 0.92. In case of NB, the feature values are conditional probabilities. So, irrespective of the range of the features, all the features are converted to the same range of 0 to 1. For this reason, NB performance is not affected as much as SVM performance when using length versus $\log_{10}$ of length.

5.4 Automatic Structuring

Model A and B outperformed the baseline model in predicting the global classes of automatic structuring. The table comparing the performance of these models can be found in the paper in part 1.

Table 5.5 shows the performance of the classes at the first level for the three models. NF ad BC are predicted better than PF by all the models. Baseline model performs much worse than model A and B in predicting PF class. CA-1 and CB-1 for both the models are same, so their performance for the 4 classes are also same.

TABLE 5.5: Prediction of first level classes in terms of F_1 score for the 3 models of automatic structuring

Classes	Baseline	Model A	Model B
PF	0.49	0.87	0.87
NF	0.94	0.95	0.95
BC	0.89	0.94	0.94
O	0.78	0.86	0.86

Table 5.6 shows the performance of the classes at the second level for the 3 models. All the sub classes of PF were predicted poorly in comparison with the sub classes of NF. The PF sub classes were better predicted by the hierarchical models than by the baseline model. NF sub classes were predicted well enough by all the 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 a 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 a aggregated location classifier CB-5.

TABLE 5.6: Prediction of second level classes in terms of F_1 score for the 3 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

Table 5.7 shows the performance of the global classes for the 3 models along with their instances. #Reports column stands for the number of reports consisting of a class. #Phrases column shows the number of phrases of each class and a phrase starts at a B-X and ends at a I-X, which precedes the start of another phrase B-X_i. #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 phrase of ‘O’ is not important, that is why there is no entry for phrases for class ‘O’. All F_1 scores are token level scores.

TABLE 5.7: 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	4229	1016	-	0.71	0.78	0.78

As can be seen from Table 5.7, 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. A part of this table is already in the paper and the important findings have already been explained. So, it will not be repeated here. This table contains some extra classes from the table in the paper and their F_1 scores have the same interpretation.

Figures 5.3, 5.4, 5.5 shows the normalized confusion matrix heat map of global classes for baseline model, Model A and Model B respectively. In baseline model (Figure 5.3), it can be seen that most classes were misclassified as other class and only BC and most NF classes were classified correctly. Some other noteworthy misclassification are 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. So, size of calcification (C) and asymmetry (AS) were misclassified with size of mass, as there were more instances of PF/MS/SI than PF/C/SI and PF/AS/SI.

For Model A and B (Figures 5.4, 5.5), there were not many misclassification with other class as 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. Model B had more true positives than Model A and baseline model had the lowest number of true positives.

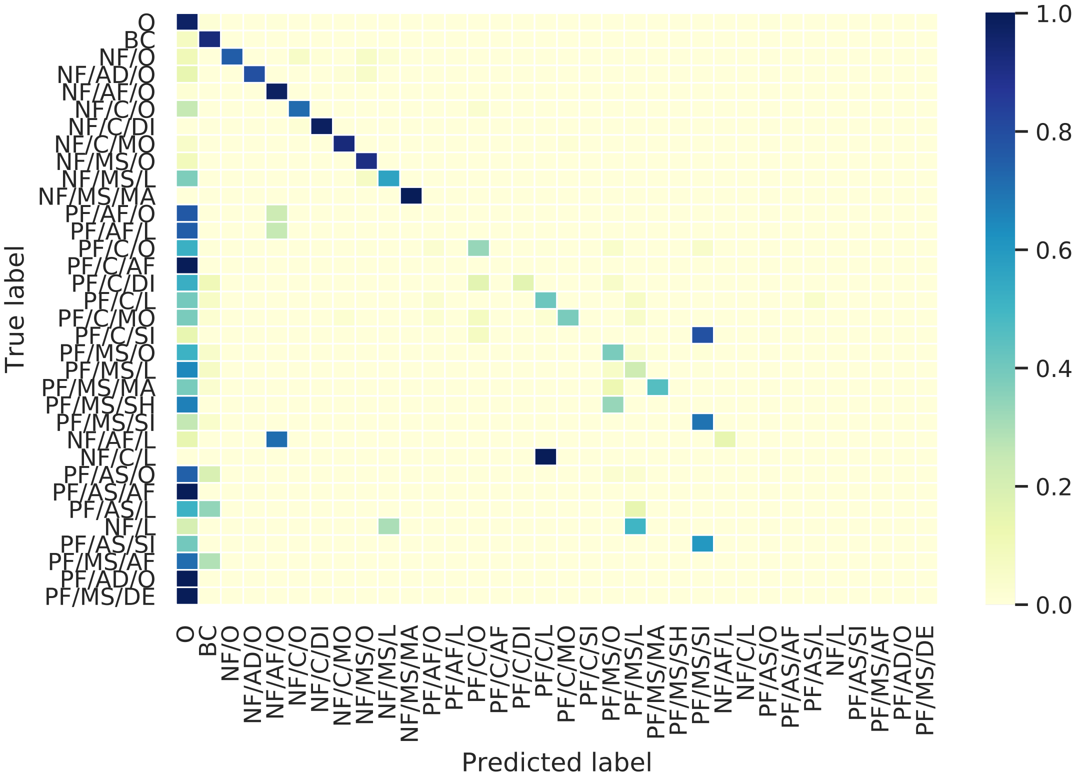


FIGURE 5.3: Normalized confusion matrix heat map: Automatic structuring baseline model

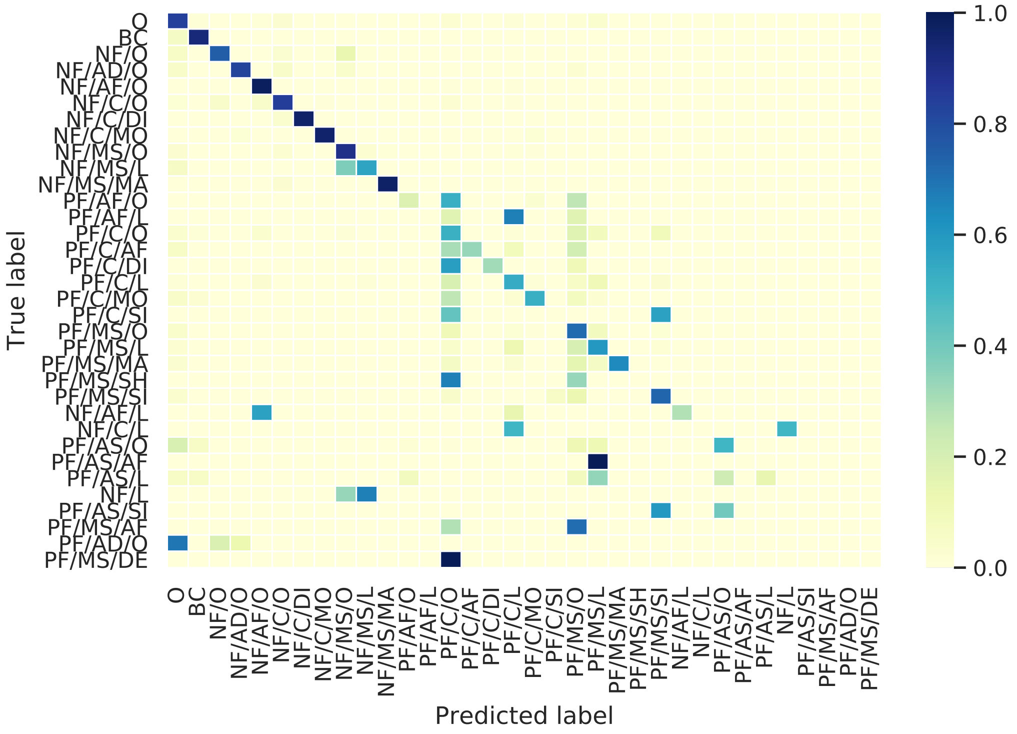


FIGURE 5.4: Normalized confusion matrix heat map: Automatic structuring Model A

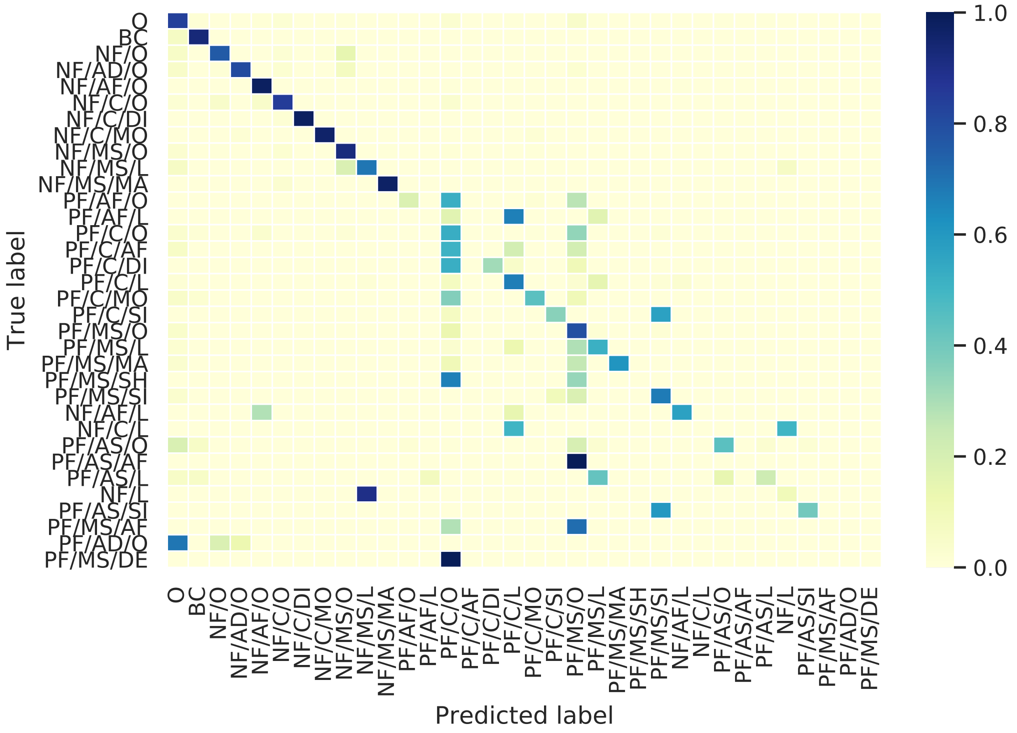


FIGURE 5.5: Normalized confusion matrix heat map: Automatic structuring Model B

Confusion matrix of Model A (Figure 5.4) and B (Figure 5.5) are almost similar. Model A and B show high false negative for PF/C/O and PF/MS/O and the classes which contribute to their high false negative value are PF/C, PF/AF, PF/AS and PF/MS. PF/MS and PF/C have the maximum number of instances (483 and 311 respectively) among all the sub classes of PF. Therefore, whenever tokens of the sub classes of PF are overlapping or not unique, they get predicted as PF/C/O and PF/MS/O. Some other noteworthy observations between Model A and B are in Model A, more NF/AF/L were getting misclassified as NF/AF/O than getting correctly classified, but in Model B, there were more true positives of NF/AF/L. Also, Model A did not have any true positive of NF/L whereas Model B had some. Similarly, PF/C/SI and PF/AS/SI had true positives in Model B. But Model A, did not have any true positives for these classes and it misclassified PF/AS/SI as PF/AS/O. This shows that there was misclassification at the third level for PF/AS. These observations prove that for Model B, aggregated classifiers of location and size helped in better prediction of location and size at the third level.

On the other hand, aggregated classifier of AF did not work at all. PF/C/AF has some true positive in Model A but none in Model B. This is because aggregated classifier for AF lost some information about its context, as the aggregated AF classifier only had other class and AF class in its surroundings. Whereas, in Model A, AF class are surrounded by distribution, morphology, location etc. during prediction, giving more insight into the surrounding of AF class, leading to better prediction.

TABLE 5.8: Performance of the individual classifiers of Model A and B in terms of F_1^M

(A) Model A

Classifiers	F_1^M
CA-1	0.90
CA-2	0.70
CA-3	0.94
CA-4	0.76
CA-5	0.72
CA-6	-
CA-7	0.91
CA-8	0.84
CA-9	0.94
CA-10	0.98
CA-11	-
CA-12	0.98
CA-13	-

(B) Model B

Classifiers	F_1^M
CB-1	0.90
CB-2	0.70
CB-3	0.94
CB-4	0.98*
CB-5	0.95
CB-6	0.97
CB-7	0.95
CB-8	0.94*
CB-9	0.97

Table 5.8 gives an overview of the performance of the individual classifiers for Model A and B. These F_1^M scores are scores on the predicted classes when true classes from previous level are given. CA-1 and CA-2 of Model A are similar to CB-1 and CB-2 of Model B, which is why they have same performance. As it can be seen, the individual classifier performance for all the classifiers is quite good. The * mark in CB-4 and CB-8 says that these scores are mainly based on 'other' class prediction by these two classifiers. Shape and density failed to get predicted by CB-4 as there were not many instances of shape and density. CB-8 failed to predict associated features

and predicted only the other class. Thus, CB-8, the aggregated classifier for associated features of PF/MS, PF/C, PF/AS, PF/AD was not a successful aggregated classifier, as these AF classes were surrounded by other classes. Model A did better prediction for AF class as Model A classifiers had information about their context e.g. AF class in PF/MS had location, size, margin, density and shape as preceding class.

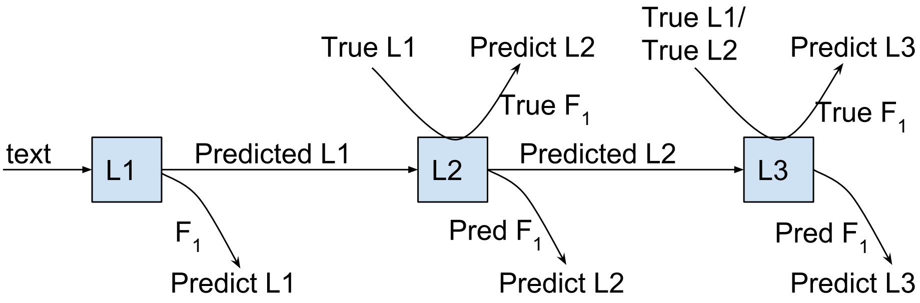
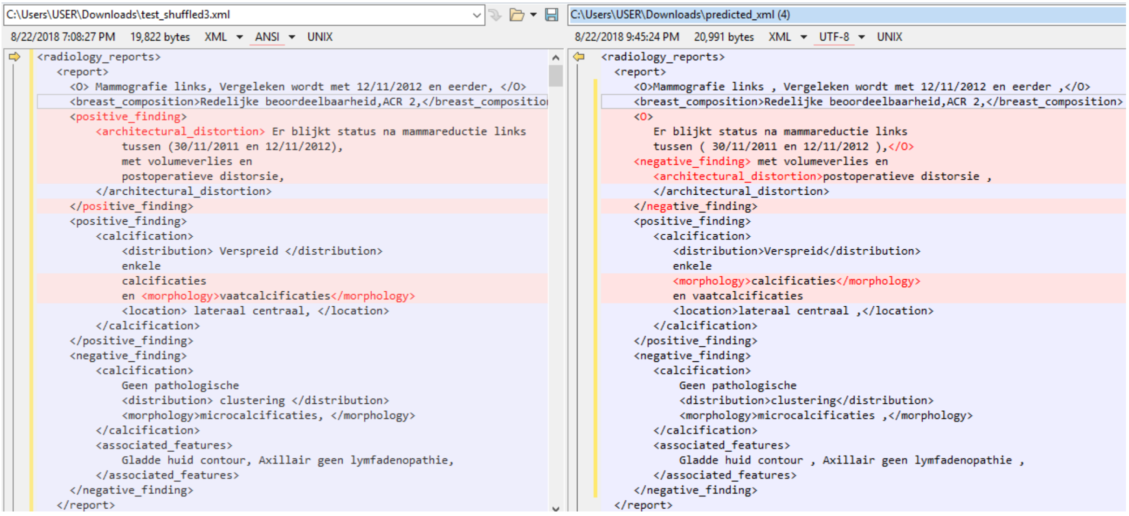


FIGURE 5.6: Error propagation through the classifiers at the 3 levels

Figure 5.6 shows the error propagation through the classifiers at the 3 levels. L1, L2 and L3 stands for the classifiers the first, second and third level. Pred F_1 at L2 is the F_1 score on predicted classes by L2 classifiers when given predicted classes from L1. True F_1 at L2 is the F_1 score on predicted classes at L2 when given true classes from L1. Similarly for L3, Pred F_1 is the F_1 score on predicted classes by L3 classifiers when given predicted classes from L2 and L1 and True F_1 is the F_1 score on predicted classes at L3 when given true classes from L1 and L2. The results corresponding to this diagram are given in the Table V in the research paper.

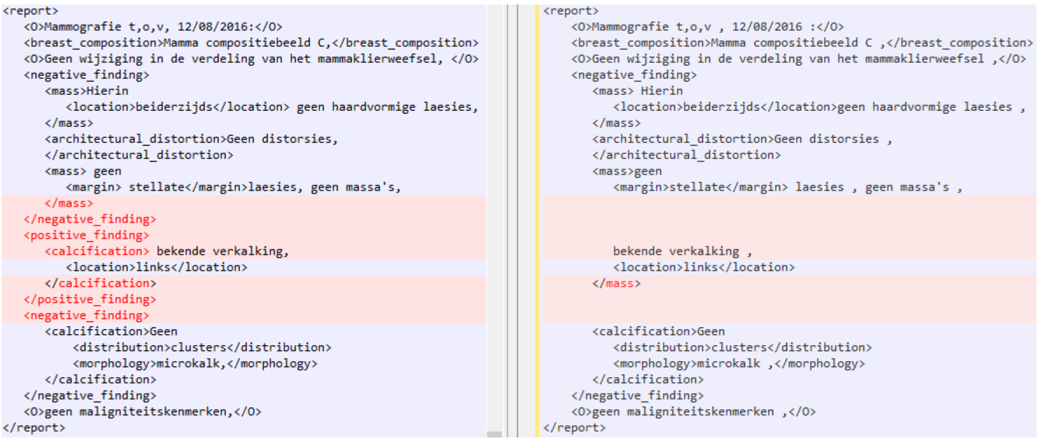

C:\Users\USER\Downloads\test_shuffled3.xml
8/22/2018 7:08:27 PM 19,822 bytes XML ANSI UNIX
C:\Users\USER\Downloads\predicted_xml (4)
8/22/2018 9:45:24 PM 20,991 bytes XML UTF-8 UNIX

```
<radiology_reports>
  <report>
    <O> Mammografie links, Vergeleken wordt met 12/11/2012 en eerder, </O>
    <breast_composition>Redelijke beoordeelbaarheid,ACR 2,</breast_compositio
    <positive_finding>
        <architectural_distortion> Er blijkt status na mammareductie links
            tussen (30/11/2011 en 12/11/2012),
            met volumeverlies en
            postoperatieve distorsie,
        </architectural_distortion>
    </positive_finding>
    <positive_finding>
        <calcification>
            <distribution> Verspreid </distribution>
            enkele
            calcificaties
            en <morphology>vaatcalcificaties</morphology>
            <location> lateraal centraal, </location>
        </calcification>
    </positive_finding>
    <negative_finding>
        <calcification>
            Geen pathologische
            <distribution> clustering </distribution>
            <morphology>microcalcificaties, </morphology>
        </calcification>
        <associated_features>
            Gladde huid contour, Axillair geen lymfadenopathie,
        </associated_features>
    </negative_finding>
  </report>
```

```
<radiology_reports>
  <report>
     <O>Mammografie links , Vergeleken wordt met 12/11/2012 en eerder ,</O>
     <breast_composition>Redelijke beoordeelbaarheid,ACR 2,</breast_composition>
     <O>
        Er blijkt status na mammareductie links
        tussen ( 30/11/2011 en 12/11/2012 ),</O>
     <negative_finding> met volumeverlies en
        <architectural_distortion>postoperatieve distorsie ,
        </architectural_distortion>
     </negative_finding>
     <positive_finding>
        <calcification>
           <distribution>Verspreid</distribution>
           enkele
           <morphology>calcificaties</morphology>
           en vaatcalcificaties
           <location>lateraal centraal ,</location>
        </calcification>
     </positive_finding>
     <negative_finding>
        <calcification>
            Geen pathologische
            <distribution>clustering</distribution>
            <morphology>microcalcificaties ,</morphology>
        </calcification>
        <associated_features>
            Gladde huid contour , Axillair geen lymfadenopathie ,
        </associated_features>
     </negative_finding>
  </report>
```

FIGURE 5.7: Automatic structuring: Comparison of the ground truth and the predicted labels by Model B of a sample report

Figure 5.7 shows comparison between the ground truth and predicted labels of a sample report for the task of automatic structuring. A comparison software called Beyond Compare [2] was used to compare the two XML reports. Figure 5.7 is a screenshot of that software showing comparison of a sample report. The left column shows the ground truth labeled report in XML format and right column shows the predicted labeled report by the Model B classifier in XML format. The pink coloured lines highlight the mismatch between the two reports and the blue lines highlight the matches. This figure gives a small idea of the mistakes that the classifier makes. Similarly, Figure 5.8 also shows the same thing for another sample report. Here, most of tokens were correctly predicted as can be seen. Only one positive finding between the two negative findings got misclassified and combined with the negative finding.



```
<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>
```

FIGURE 5.8: Automatic structuring: Comparison of the ground truth and the predicted labels by Model B of another sample report

Chapter 6

Conclusion

In this section, at first, we discuss to what degree our research questions mentioned in Chapter 1 were answered and how. The main research question is explained followed by the three sub research questions. There is a discussion section after this which summarizes our work and the findings, and also lists the limitations of our work. This chapter ends with a section on future work.

6.1 Research Questions

Main RQ: *To what degree can we successfully conduct quality assurance of radiology reports using machine learning?*

ACR BI-RADS has set some protocols for radiology reporting on breast cancer. For the purpose of quality assurance, we developed a scheme related to ACR BI-RADS to check if the reports produced by the hospitals conform to the well-defined protocols. We successfully built machine learning models to detect the presence of the reporting structure designed by ACR BI-RADS. Our models can detect the presence of indication of examination (clinical data), breast composition, clear description of findings and conclusion. We created a semi-structured XML format for the findings such that important findings can be clearly visualized. This semi-structured format can also be used for other purposes like easy extraction of information and increased readability of reports for clinicians. Our models were not used for actual trials in hospitals to check how many reports are actually conforming to protocols. This can be done through development of a prototype of our models. Our models only check for the presence of the important entities of a report and not for the accuracy of the content. Thus, quality assurance done in our project can help in assessing report clarity and organization. Report accuracy can also be partially assessed, but only from the perspective of the existence of a content and not its verification.

Sub RQ-1: *How can we identify the most apparent top level structure from the report using machine learning?*

The top level structure, most apparent from the report, involves identification of the headings in the report. We predicted the heading, not heading and title classes of a report with an F_1 score of 0.98, 0.98 and 0.98 respectively using SVM. The names of the headings provide with the information about the important top level sections in the report.

Sub RQ-2: *How can we automatically verify if the information in the report has been placed under the correct top level sections?*

The top level sections in a report are clinical data, findings and conclusion. We predicted these sections with an F_1 score of 0.94, 0.94 and 0.92 respectively using SVM. These are also the names of the headings that can found in Sub RQ-1. If a post-processing step is executed, then the sentence predicted as heading in Sub RQ-1 can be matched with the predicted section of Sub RQ-2. If these two information match, then we can say that the information in the report has been placed under the correct top level sections (or precisely headings).

Sub RQ-3: *To what extent can we automatically convert the free-text findings from the report into a detailed structured format?*

We build 3 CRF models for converting the free-text findings from mammography study to semi-structured XML format. Our hierarchical models, Model A and B outperformed our baseline model with an F_1 score of 0.78 vs 0.71. Breast composition, positive finding and negative finding were predicted with an F_1 score of 0.94, 0.87 and 0.95 respectively by Model A and B. Mass and calcification were better predicted than asymmetry and associated features in positive finding. In negative finding, all the classes were predicted with an F_1 score>0.88 and associated features was predicted the best with an F_1 score of 0.96.

6.2 Discussion

We developed a method for automatic structuring of Dutch free-text radiology reports on breast cancer for quality assurance using machine learning algorithms. We divided the method into 3 steps – i) identification of the sentences in the report as headings, not headings and title (Heading identification), ii)identification of the content of the report as clinical data, findings and content (content identification), and iii)finding a detailed structure of the mammography findings in the report and automatically converting the unstructured mammography findings to a semi-structured format (automatic structuring). The first two tasks used simple machine learning classification algorithms like Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF). The third task used an classification algorithm for sequence labeling called Linear Chain Conditional Random Field (LC-CRF).

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 but improved F_1 score of RF classifier (but did not result in a better score than SVM) and lowered F_1 score of NB 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 hierarchical CRF models (Model A and B) outperformed the baseline CRF model with F_1 score of 0.78 vs 0.71.

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 had very specific way of describing themselves unlike positive finding, which used varied vocabulary to describe the findings. This made the prediction for positive finding harder than the other two.

Second level positive finding classes – mass and calcification were predicted better than asymmetry, associated features and architectural distortion. This is because far lesser training data was available for the latter classes than the former ones. Also, on discussion with radiologists, it was understood that asymmetry findings are always hard to understand. So, low score 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 the specific word cluster and margin can be described using specific 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 less training data (around 2 or 3 tokens). Both second level and third level sub classes of associated features in positive finding was 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. Example of these classes are 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, each token are surrounded by various other classes in that group, for example, 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. Some examples of these classes are margin, morphology, distribution and associated features at third level under positive finding.

In our baseline model, misclassification mainly consisted of non-other classes getting wrongly predicted as other class. Similarly, 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 high level. For example, positive finding/calcification/distribution get 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. Poor quality reports get predicted as poor quality, as other class is never wrongly predicted as a non-other class by our models. So, for the purpose of quality assurance, our aim to identify the poor quality reports can be solved by our models.

Though, an analysis of how many reports conform to reporting standards was not done for the scope of this project, a table in experiments chapter (Table 5.7) was provided showing the number of reports containing each class. This table showed that shape and density of the mass existed in only 3 reports and 1 report respectively out of 108 mammography reports. This shows that these were the two classes least reported in the findings (a similar thing was also reported by Houssami et al. [13] in 2004). Whether this was a mistake from the point of view of reporting or these 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 is a lot more than only 24% reports containing breast composition as reported by Houssami et al. [13]. But according to ACR BI-RADS guidelines, all reports should contain a breast composition. These type of analysis can extended for quality assurance as a future work.

The automatic structuring models developed in this thesis can help in making the information in the reports searchable. The huge volume of information can be harvested and used for other research purposes, for example, for answering questions, like, how many patients had lesions in their right breast? It will also become easier for 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.

To the best of our knowledge, the work done in this thesis has not been done before. A similar work was done by Esuli et al. [7], for information extraction from mammography findings written 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 [10] applied CRF for information extraction in chest CT radiology reports written in English. They had a comparable number of reports to ours, i.e. 150 reports and 5 classes in their model. They used 10 fold cross validation on 150 labeled reports, achieving a F_1 score of 85.3%. We can say that though 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, with much less training data. Increasing training might increase the performance of our models as well.

6.3 Limitations

This project has some limitations which are discussed next. The major limitation is that no labeled dataset was already available for the task of automatic structuring and they needed to be labeled manually by the radiologists at ZGT. The labeling scheme developed for this purpose was extensive and time consuming and it is hard for experts, such as radiologists, to find time to label huge number of reports for training and validation of our models. So, due to time constraints, only 108 reports could be labeled for building our model and the results shown in the thesis are for 4 fold cross validation, which means 27 reports as test set in each fold. This is a very small number of test set and the results will vary if more labeled data set is available. This small number of labeled reports may not be the most perfect representation of all the reports on breast cancer produced by the radiologists at ZGT. Also, there were no training data available for 5 classes out of 39 classes in our model – positive finding/architectural distortion/location, positive finding/architectural distortion/associated features, negative finding/architectural distortion/location, negative finding/asymmetry/location and negative finding/asymmetry/other, and thus these classes could not be trained.

For the task of automatic structuring, we randomly shuffled the mammography findings. F_1 scores in each fold did not vary much for the combination of shuffled data used for this project. We also tested on other combinations of shuffled data, where the results in each fold varied with lowest around 0.68 and highest around 0.82. The shuffled data used in this project was chosen to represent results when all folds have around similar balance of classes.

Another limitation is that for automatic structuring, we only focused on findings from mammography study. The findings section is the radiology report contains findings from mammography study, ultrasound and MRI and findings from each of these study have different structure. Training the system on all the types of findings would mean developing a structure for each of the study and manually labeling findings from each of the study by following the structure. Due to time constraint, we only developed a structure for mammography study and trained models on it, but the similar models like ours can be developed for other studies as well.

Another limitation is that the findings from the mammography study were manually extracted from the radiology report. It will be easier if this process could be automated by training the system to recognize the mammography findings. More about this is explained in future work. Our project focuses on radiology reports on breast cancer. Similar models can also be extended on radiology reports about other conditions. Also, we focused on Dutch radiology reports and similar models can be extended for other languages.

Another limitation is that two radiologists collaboratively annotated 18 reports and separately annotated 45 reports each. Though, annotations were checked by a trained expert to remove any inter-annotator discrepancies but some discrepancies may still remain.

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, that gets propagated to the next level and makes the rest of the prediction by second and third level classifiers wrong. Our models do not contain a way to mitigate the error propagation. One way of how this could be handled is discussed in future work.

Our project automatically structures the reports for the aim of quality assurance but does not check the reports for quality assurance as an end result. The project can be extended to check for quality in the reports produced by the ZGT.

6.4 Future Work

There are several possibilities that the research can be extended to. Some of these are described in this section.

The three steps described above in the discussion section 6.2 are separate steps independent of each other. These steps can be linked to each other and the whole process can be made a automatic one. The headings identified at the first step, can be taken as the top level sections of the report. These headings contain the names, clinical data, findings and conclusion. In the second task, the clinical data, findings and conclusion can be predicted. Then, the headings of the first step can be compared with the content (section) identification of the second task. If the name of the identified headings match with the correct content, then it can be said that the headings correspond to the correct sections and the report was written correctly.

For the content identification, instead of identifying only three classes – clinical data, findings and conclusion, more classes can be predicted. For example, findings can be divided into mammography, ultrasound and MRI. Conclusion can be separated into conclusion and BI-RADS category. In summary, the classes to be predicted from the whole report can be clinical data, mammography findings, ultrasound findings, MRI findings, BI-RADS category and conclusion. This will help for the third step i.e. automatic structuring. The section identified as mammography can be taken as an input for the third task of automatic structuring. Then, no manual extraction will be needed, assuming that section identification results in a near perfect accuracy.

Another possibility is the whole task of content identification can be seen as a sequence labeling problem and Conditional Random Field (CRF) can be used for predicting the sequence of the classes mentioned in previous paragraph. Automatic structuring also can be extended for ultrasound by following the structure given by ACR BI-RADS.

All the things described above can be combined to develop a prototype for quality assurance and this can be sent for clinical trial. This will help in checking how many good quality reports a hospital is producing. If a report is not of the expected quality, then the concerned radiologist can be asked to improve it. This will help in monitoring the clinical performance of a hospital. Hospitals have a check list to monitor if the reports produced by the hospital contain the items in the checklist and checking the reports and filling up the checklist is usually done manually. According to the radiologists, automated way of filling up of checklist is unheard of. If our models can be implemented in actual clinical scenario, then quality assurance of reports and further filling up of checklist can be automated. This will be very helpful for the hospitals.

Another possibility is checking the presence of BI-RADS category and also predicting it. ACR BIRADS states that a report must contain a BIRAD category indicating the malignancy of the breast cancer. A BIRAD category can take a value from 0-6, 0 being benign and 6 being most malignant. The presence of this category is seen in the conclusion section. For the purpose of quality assurance of reports, the check that a report contains a BI-RADS category, is a must. Based on the findings section listed in the report, the BIRAD category can also be predicted to see how well it matches with the category assigned by the radiologists.

Further, the BI-RADS category assigned can be checked with the pathology reports, which is the ultimate decision maker. If something gets confirmed through pathology reports, then that is taken as the true result. On comparing the assigned BI-RADS category with pathology outcome, it can be checked that how well the radiologists are doing the job of assigning a BIRAD category to the reports.

The models used in the project can also be improved such that their performance increases. One method can be adding a positional feature to our models. The position of a token in the report can contain valuable information in predicting a class more accurately. For example, clinical data is always written at the beginning of the report, followed by the findings and then the conclusion. Similarly, breast composition comes at the beginning of the mammography findings, then the positive or negative finding. Thus, adding a feature, containing the position of the token, can improve the accuracy. Esuli et al. [7] experimented with positional feature for the task of information extraction from Italian radiology reports on breast cancer, using CRF but they concluded that positional feature did not bring any substantial difference in the the performance of the model. Therefore, this is something that can be experimented with.

Model A and B in our model has their own advantages and disadvantages as explained in the discussion section 6.2. The concept behind these two models can be combined into one to make a model with advantages of both. Model A can predict the classes by following the suggested hierarchical structure of the report. Aggregated classifiers can also be used to predict classes as done in Model B. Then, for each token, the probabilities of class prediction by both the models can be compared and class with maximum conditional probability can be assigned to the token.

Another idea to improve prediction in our models is checking the class with the second highest conditional probability. In the automatic structuring problem, it was observed that many times, the correct class was the one with second highest conditional probability. So, the idea is to pass both the highest probability class and second highest probability class to the classifiers at the next level. For example, suppose at the first level, a token has positive finding as the highest probability class and negative finding as the second highest probability class at the first level, then the token will be passed to both positive finding and negative finding classifier at the second level for further prediction. At the end of passing through all the levels, the overall probabilities of the global class at the 3 levels will be averaged for both the best and second best predicted class and the combination with the highest probability will be assigned to the token as its global class.

A possible problem in this model is error propagation through all the levels. This can be addressed by using Factorial Conditional Random Field, a type of dynamic CRF (or general CRF), in place of Linear Chain CRF. Factorial CRF is a combination of multiple LC-CRFs, one for each output level. Factorial CRF jointly predicts classes at all levels of hierarchy, thus helping to address the issue of error propagation.

Another analysis that can be done is related to inter annotator discrepancies. As there were two radiologists annotating the reports separately, how well do their annotation match with each other can be analyzed. This can help in comparing human error and machine made error. This idea was also implemented in Esuli et al. [7] and accuracy measure by the systems were found to be higher than the inter annotator agreement.

Similar models developed in this project can be extended to reports on conditions other than breast cancer and written in different languages. It will be interesting to see how the models perform in this scenario.

Bibliography

[1] RR Armas. "Qualities of a good radiology report." In: *AJR. American journal of roentgenology* 170.4 (1998), pp. 1110–1110.

[2] *Beyond Compare*. `https://www.scootersoftware.com/`.

[3] Brian E Chapman et al. "Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm". In: *Journal of biomedical informatics* 44.5 (2011), pp. 728–737.

[4] Matthew C Chen et al. "Deep learning to classify radiology free-text reports". In: *Radiology* (2017), p. 171115.

[5] Koby Crammer and Yoram Singer. "On the algorithmic implementation of multiclass kernel-based vector machines". In: *Journal of machine learning research* 2.Dec (2001), pp. 265–292.

[6] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. "What can natural language processing do for clinical decision support?" In: *Journal of biomedical informatics* 42.5 (2009), pp. 760–772.

[7] Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. "An enhanced CRFs-based system for information extraction from radiology reports". In: *Journal of biomedical informatics* 46.3 (2013), pp. 425–435.

[8] Carol Friedman et al. "A general natural-language text processor for clinical radiology". In: *Journal of the American Medical Informatics Association* 1.2 (1994), pp. 161–174.

[9] Carol Friedman et al. "Natural language processing in an operational clinical information system". In: *Natural Language Engineering* 1.1 (1995), pp. 83–108.

[10] Saeed Hassanpour and Curtis P Langlotz. "Information extraction from multi-institutional radiology reports". In: *Artificial intelligence in medicine* 66 (2016), pp. 29–39.

[11] Marti A Hearst. "Untangling text data mining". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics. 1999, pp. 3–10.

[12] PM Hickey. "Standardization of roentgen-ray reports". In: *AJR Am J Roentgenol* 9 (1922), pp. 422–425.

[13] Nehmat Houssami et al. "Quality of breast imaging reports falls short of recommended standards". In: *The Breast* 16.3 (2007), pp. 271–279.

[14] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991* (2015).

[15] Annette J Johnson et al. "Cohort study of structured reporting compared with conventional dictation". In: *Radiology* 253.1 (2009), pp. 74–80.

[16] Annette J Johnson et al. "Improving the quality of radiology reporting: a physician survey to define the target". In: *Journal of the American College of Radiology* 1.7 (2004), pp. 497–505.

[17] Charles E Kahn Jr et al. “Toward best practices in radiology reporting”. In: *Radiology* 252.3 (2009), pp. 852–856.

[18] John Lafferty, Andrew McCallum, and Fernando CN Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In: (2001).

[19] Curtis P Langlotz. *RadLex: a new method for indexing online educational materials*. 2006.

[20] Dingcheng Li, Karin Kipper-Schuler, and Guergana Savova. “Conditional random fields and support vector machines for disorder named entity recognition in clinical texts”. In: *Proceedings of the workshop on current trends in biomedical natural language processing*. Association for Computational Linguistics. 2008, pp. 94–95.

[21] Houssam Nassif et al. “Information extraction for clinical data mining: a mammography case study”. In: *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*. IEEE. 2009, pp. 37–42.

[22] Ewoud Pons et al. “Natural language processing in radiology: a systematic review”. In: *Radiology* 279.2 (2016), pp. 329–343.

[23] Felicity Pool and Stacy Goergen. “Quality of the written radiology report: a review of the literature”. In: *Journal of the American College of Radiology* 7.8 (2010), pp. 634–643.

[24] Daniel K Powell and James E Silberzweig. “State of structured reporting in radiology, a survey”. In: *Academic radiology* 22.2 (2015), pp. 226–233.

[25] Bruce I Reiner, Nancy Knight, and Eliot L Siegel. “Radiology reporting, past, present, and future: the radiologist's perspective”. In: *Journal of the American College of Radiology* 4.5 (2007), pp. 313–319.

[26] Lee Robert, Mervyn D Cohen, and Greg S Jennings. “A new method of evaluating the quality of radiology reports”. In: *Academic radiology* 13.2 (2006), pp. 241–248.

[27] WC RONTGEN. “Ueber eine neue Art von Strahlen (2. Mitteilung)”. In: *Sitzungsberichte der Phikalisch-medizinischen Gesellschaft zu Wurzburg* 2 (1896), pp. 11–17.

[28] Lawrence H Schwartz et al. “Improving communication of diagnostic radiology findings through structured reporting”. In: *Radiology* 260.1 (2011), pp. 174–181.

[29] *SEER Training Modules, Quadrants of the Breast. U. S. National Institutes of Health, National Cancer Institute.* https://training.seer.cancer.gov/breast/anatomy/quadrants.html. (Visited on 08/21/2018).

[30] E. A. Sickles et al. “ACR BI-RADS® Mammography. In”. In: *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA, 2013.

[31] Dorothy A Sippo et al. “Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing”. In: *Journal of digital imaging* 26.5 (2013), pp. 989–994.

[32] Chris L Sistrom and Janice Honeyman-Buck. “Free text versus structured format: information transfer efficiency of radiology reports”. In: *American Journal of Roentgenology* 185.3 (2005), pp. 804–812.

[33] Chris L Sistrom and Curtis P Langlotz. “A framework for improving radiology reporting”. In: *Journal of the American College of Radiology* 2.2 (2005), pp. 159–167.

[34] *Support Vector Machine*. https://en.wikipedia.org/wiki/Support_vector_machine. (Visited on 08/22/2018).

[35] Charles Sutton, Andrew McCallum, et al. "An introduction to conditional random fields". In: *Foundations and Trends® in Machine Learning* 4.4 (2012), pp. 267–373.

[36] Joost Timmerman et al. "Automatically structuring free-text radiology reports using a machine learning algorithm". In: (2014).

[37] Manabu Torii, Kavishwar Wagholikar, and Hongfang Liu. "Using machine learning for concept extraction on clinical documents from multiple data sources". In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 580–587.

[38] Shijun Wang and Ronald M Summers. "Machine learning and radiology". In: *Medical image analysis* 16.5 (2012), pp. 933–951.

[39] Jeffrey B Ware et al. "Effective radiology reporting". In: *Journal of the American College of Radiology* 14.6 (2017), pp. 838–839.

[40] Meliha Yetisgen-Yildiz et al. "A text processing pipeline to extract recommendations from radiology reports". In: *Journal of biomedical informatics* 46.2 (2013), pp. 354–362.

[41] Meliha Yetisgen-Yildiz et al. "Automatic identification of critical follow-up recommendation sentences in radiology reports". In: *AMIA Annual Symposium Proceedings*. Vol. 2011. American Medical Informatics Association. 2011, p. 1593.

[42] Harry Zhang. "The optimality of naive Bayes". In: *AA* 1.2 (2004), p. 3.

[43] Shaodian Zhang and Noémie Elhadad. "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts". In: *Journal of biomedical informatics* 46.6 (2013), pp. 1088–1098.

[44] David Zingmond and Leslie A Lenert. "Monitoring free-text data using medical language processing". In: *Computers and Biomedical Research* 26.5 (1993), pp. 467–481.

[45] Harmien Zonderland and Robin Smithuis. *Radiologyassistant.nl. (2013). Bi-RADS for Mammography and Ultrasound 2013*. http://www.radiologyassistant.nl/en/p53b4082c92130/bi-rads-for-mammography-and-ultrasound-2013.html.