UNIVERSITY OF TWENTE.

Interaction Technology

An Interactive NLP Approach for Improving Completeness and Annotation Efficiency in Prostate Screening Reports

Hridya Nair Suresh
3096874
Graduation Project

Supervisors: Dr. Shenghui Wang
Dr. Doina Bucur
Jeroen Geerdink

Faculty of Electrical Engineering,
Mathematics and Computer Science(EEMCS)
Human Media Interaction (HMI)
University of Twente
Ziekenhuisgroep Twente
The Netherlands

Summary

Radiology is an important component of healthcare, playing a vital role in disease diagnosis. The process of radiology reporting, where radiologists document their findings and observations from scans, is integral to patient care. Thus, the completeness of these reports is essential, as minor errors can significantly affect the diagnosis and further treatment. The mistakes or missing fields in the report can arise due to factors such as increased workload, time constraints and inexperienced radiologists. This research focuses on automating the process of checking reports and providing radiologists with suggestions for any missing information.

The radiology reports are Dutch semi-structured text data, Natural language Processing(NLP) techniques were used to extract the important information from the reports. Dutch Language models BERTje and MedRoBERTa.nl were tested for this task, but they exhibited overfitting due to a limited dataset. A hybrid Conditional Random Field (CRF) model was also implemented, yielding better results with an F1 score ranging from 0.94 (highest) to 0.45 (lowest) in identifying fields. The lower performance for certain labels is attributed to the underrepresentation of these fields in the reports.

To address the challenges of limited data and underrepresentation, we developed an interface that integrates the model into the radiologists' workflow, allowing for both the application of the model and the collection of annotations through user interactions.

Acknowledgement

I would like to express my sincere thanks to my supervisors Shenghui Wang, Jeroen Geerdink and Doina Bucur, for their valuable feedback and support throughout this journey. I am also grateful to the radiologists Jeroen Veltman and Lars Bosboom from ZGT for dedicating their valuable time to participate in the evaluation process and a special thanks to Jeroen for coordinating these sessions. Last, but not the least, I would like to thank my family and friends whose support and encouragement helped me power through the challenging days of my thesis. I am truly thankful for all of your support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Radiology reports play a crucial role in diagnostic decision-making and treatment planning in healthcare. The clarity and completeness of radiology reports are important, as they are the primary means of communication between radiologists, referring physicians and other healthcare workers involved in patient care [1] [2]. However, despite the significance, ensuring their completeness remains a considerable challenge in the field of radiology [3] [4]. This is due to several factors such as high workload, time constraints, and lack of experience [5] [6]. Automating the checking of radiology reports can be an approach to address this challenge, and support radiologists in their work.

Ensuring that radiology reports are both complete and compliant with established guidelines is vital for accurate diagnosis and effective treatment. Usually, radiology reports have a general structure with patient data, findings, discussion, and recommendations [7]. To maintain the high quality of radiology reports, the hospitals issue guideline documents that outline the expected structure and essential components of these reports [8].

An incomplete radiology report may lack the necessary fields and corresponding values, and a non-compliant report may contain all required fields but fail to adhere to the structure specified in the relevant guideline documents. Such incomplete or non-compliant reports can compromise the accuracy of the report and, consequently, the patient's treatment. In most cases, rechecking the report manually is the approach, but it is time-consuming and adds to the already substantial workload of radiologists. To address this challenge, one of the focuses of our research is to automate the checking of radiology reports, aiming to assist radiologists in ensuring their reports are both complete and compliant. The automation of checking the reports involves identifying missing elements in the report and providing them as suggestions to the radiologists.

This research is conducted in collaboration with Ziekenhuisgroep Twente (ZGT). The prostate screening reports and the guideline document provided by the hospital are used as the dataset for our study. To develop the model, we reviewed previous literature where NLP has been used in information extraction tasks. In recent years, Natural Language Processing (NLP) has emerged as a powerful aid in the field of radiology [9] [10] [11]. NLP models have been employed for extracting essential information from radiology reports and performing tasks like pattern matching, which has led to the automation of various processes within the radiology workflow such as Clinical Decision Support (CDS), summarization, and quality control. Given that our dataset comprises semi-structured text-based radiology reports, and our task involves identifying key fields within these reports, NLP is considered an effective approach for enhancing their completeness and compliance. This leads to our first research question with the following sub-questions:

”How accurately can NLP algorithms cross-check prostate screening reports with Guidelines to improve their completeness and compliance?”

RQ1: *”What models can be adapted to extract information from the report?”*

RQ2:*”How can we ensure the compliance of the report?”*

RQ3:*”What evaluation metrics should be used to measure the performance of the models?”*

While developing models to assist radiologists is essential, integrating these models into their workflow is also equally important. The challenges in implementing artificial intelligence models in radiology include the uncertainty of using them in clinical applications [12]. One of the reasons is the lack of standard user interfaces to integrate the model results into their workflow [13]. Therefore, another key objective of this research is to explore how these models can be effectively integrated using an interface into the reporting process.

Additionally, we aim to use this interface to capture user interactions that can be converted into annotations, which are an integral part of training models. Our current dataset contains 206 prostate screening reports annotated by radiologists, which is a relatively small amount. This annotation is a time-consuming process, that requires the expertise of radiologists, making it difficult to obtain a large, high-quality annotated dataset. This contributes to the poor performance and overfitting in Language Models. Although the hybrid CRF model demonstrated better results, it still underperformed in fields that were underrepresented in the training data.

To address this, we propose a solution of using the user interface not only to integrate the model outputs but also to facilitate the annotation as part of the radiologists’ reporting procedure. This approach minimizes additional workload and simultaneously enhances the dataset used for training the model, thereby improving its accuracy over time. This method aligns with the concept of Human-in-the-Loop (HILT), where user interactions are leveraged to continuously refine the model, building trust and engagement among users. This brings us to our second research question and the following sub-questions:

" How can models be integrated into the reporting workflow of radiologists to check the report and to improve the annotation process simultaneously?"

RQ1: *"What features should the user interface include to support seamless integration of NLP techniques into radiologists' workflow?"*

RQ2: *"What features should the interface have, to adapt model learning from user interactions?"*

RQ3:*"What metrics should be used to evaluate the effectiveness of NLP integration improving radiologists' reporting practices and annotation process?"*

This thesis begins with a review of related literature on the application of NLP in radiology, model integration into workflows, and the Human-in-the-Loop (HILT) in Chapter 2: **Related Works**. This will be followed by Chapter 3: **Dataset** describes the dataset used in this research. Chapter 4: **Identify the Missing Fields and Ensuring Compliance of Reports** answers our first research question, while Chapter 5: **Interface and Annotation** addresses our second research question. Finally Chapter 6 provides the **Discussion**, and Chapter **??** presents the **Conclusion and Recommendations.**

Chapter 2

Related Works

This chapter discusses previous studies on the application of Natural Language Processing (NLP) in radiology along with the integration of Models and the concept of ”Human in the Loop” (HITL). These studies are particularly relevant to the challenges we are addressing.

The chapter is organised into six sections:
Section 2.1: NLP applications in Radiology
Section 2.2: NLP approaches to extract information
Section 2.3: Dutch Language Models
Section 2.4: Prompting Language models
Section 2.5: Integrating AI Models with Human-in-the-Loop Systems.

2.1 NLP applications in Radiology

Numerous studies have explored the application of NLP to improve the efficiency and accuracy of radiology reporting. This includes Clinical Decision Support (CDS), Quality control, Performance monitoring, Increasing diagnostic accuracy, Early patient Prognosis, Findings alert, Improved productivity, Reporting findings, Follow-up of Test results, and Choice of Procedures [14]. Some of the tasks in these applications are useful for our project, including text extraction, pattern matching, and entity tagging.

For instance, Nguyen et al. [15] propose a hybrid model for automating the summarization of Dutch breast cancer Radiology Reports. They combined an encoder-decoder attention (EDA) model with a BI-RADS score classifier. While the primary task was summarization, a subtask involved obtaining the BI-RADS score, an important field in breast cancer radiology reports. This shows the importance of identifying and examining key-value fields, an aspect that is considered in our research as well. Another interesting study by Shreyasi et al. [11] uses a combination of NLP and ML to structure Radiology Reports of Breast Cancer Patients for Clinical Quality Assurance. They used a hierarchical Conditional Random Field (CRF) and Support Vector Machine to achieve the automated structuring and obtained an F1 score of 0.78.

Donnelly et al. [16] reviews studies that used NLP technologies to evaluate radiology reports. This literature study explains the basics of NLP techniques and how they can be used in assessing radiology reports. Extracting information from the report is a fundamental step for most of the tasks mentioned, as the reports can be unstructured or semi-structured. They divide the approach of NLP extraction of texts into two: symbolic or rule-based and statistical or machine learning techniques. Notably, several studies have adopted a hybrid approach, combining both techniques. The next section discusses the various strategies employed in previous research to extract information from radiology reports.

2.2 NLP approaches for Information extraction

Extracting vital information from the reports is the basic step in most of the NLP tasks in radiology. This also plays an important role in our research problem, as identifying and searching for relevant fields is essential for assessing the completeness of reports and providing feedback. Several approaches have been used according to the specific requirement of extraction. One of them is Named Entity Recognition (NER).

Named entity recognition models are used to extract vital information from free texts. María et al. [17] provide a methodological literature review of NER models used in Electronic Health Records. According to this study, Deep Learning models were the most prevalent, accounting for 58.86% of the approaches. Traditional Machine Learning methods followed, constituting 20.75% of the NER techniques. Graphical models like CRF and rule-based approaches held a smaller share, representing 13.20% and 6.79% of the methods, respectively.

In radiology reports, the measurements are an important part and thus extracting them for any of the applications should be accurate. Selen et al. [18] propose a hybrid NLP pipeline that can extract measurements and descriptors from free text radiology reports. The pipeline consists of an automated NER tagging using the Condition Random Field (CRF) model and a rule-based measurement tagging using regular expression. The model was trained on 1117 reports and obtained a good performance with an F1 score ranging from 80% - 98% for the information types.

Language Models are another approach explored in other studies. BioBERT is a deep learning language model that is specifically for the medical domain [19]. It is a variation of the BERT (Bidirectional Encoder Representations from Transformers) model that is pre-trained using PubMed abstracts and PubMed Central full-text articles. Xin Yu et al. [19] proposes a BioBERT-based NER in the electronic medical record to annotate clinical problems, treatments and tests. This model was trained on a dataset consisting of 426 discharge summaries, with 170 used for training and the remaining 256 reserved for testing. The model achieved an impressive F1 score of 87.10%. In our case, the reports are in Dutch thus Dutch language models are more suitable in our case. The next section discusses Dutch LMs.

2.3 Dutch Language Models

While the previously mentioned studies primarily focused on extracting information from English medical reports, it is essential to note that Dutch reports are the subject of our current research. Therefore it is necessary to explore studies specifically dealing with Dutch medical reports.

BERT models can be modified to adapt them to specific tasks in three ways: pre-trained on a generic corpus, pre-trained on a generic corpus and further on a domain corpus, pre-trained exclusively on a domain corpus. BERT-NL, BERTje, and RobBERT are three domain-generic Dutch models that are trained on general data from Wikipedia, news and web data. To achieve a good performance in medical data, domain-specific Dutch models are necessary. MedRoBERTa.nl is the first language model for Dutch medical Records proposed by Stella et al. [20]. They used the RoBERTa as the base model and used Dutch hospital notes to pre-train the model from scratch. This model has outperformed the general model on the task of odd-one-out similarity for Dutch medical records. MedRoBERTa.nl also outperformed the general model when fine-tuned to the task of classifying sentences from Dutch hospital notes that contain information about patients' mobility levels.

Contradicting the previous study, the research by Rietberg et al. [21] shows that BERTje, the generic Dutch model has outperformed the MedRoBERTa.nl on the task of extracting the reason for taking an MRI scan of Multiple Sclerosis (MS) patients using the attached dutch free-form reports. This shows that domain-specific models are not always superior. They also noted that BERTje performed better than both RobBERT and MedRoBERTa.nl which both are RoBERTa-based, thus it could also be that BERT-based models are better for their particular task. Therefore comparing the performance of different models may help to get to the model perfect for our task.

2.4 Prompting techniques for limited annotated data

Another challenge associated with our Dataset is the availability of annotated data. To train a supervised model efficiently, good-quality annotated data is a necessity. However, the manual effort and time for this process are expensive, especially in the medical domain. In our case, the annotations are done by the radiologists and they are already busy with their job. Thus obtaining a large amount of annotated data is not ideal. Therefore methods that can perform with limited annotated data have to be explored. One such method is prompting the model.

Prompting generally doesn't require training samples, but in some cases, training data are given to the model to understand the task. The first case is known as the zero-shot setting, where the model recognises new tasks through its description [22]. The other approach is where a large amount of training data is given to the model known as full-data learning or very few samples are given known as few-shot learning. [23]. The few-shot or zero-shot approach can be preferred more in the case of limited data.

Multiple prompts can be used to train the language model to perform a task and several methods have been explored for this. This can be helpful, as more than one field has to be analysed for our research problem. Prompt Ensembling is one such approach which combines multiple unanswered prompts as input to the model. This can help combine the multiple aspects of the task. Arora et al. [24] proposed a similar prompting approach called Ask Me Anything (AMA), showing that a single question can be reformatted as different prompts. Then they combine the intermediate answers from these prompts to obtain the final output. DIVERSE is another prompt ensembling approach proposed by Yifei L, et al. [25] where they have the same method for developing prompts as AMA but use a voting verifier (a neural network) for selecting the answer as the final output.

Prompting has also been explored in performing tasks in the medical domain. [26] shows the use of the prompting technique in natural language generation of justification of medical diagnosis given the case description and disease symptomatology. Similarly prompt tuning is also used in the classification in the field of medical data [27] [28].

2.5 Integrating AI Models with Human-in-the-Loop Systems

Implementing an interface to integrate the model is a way in which we can ensure that radiologists can use the model in completing the task. The implementation should be easily accessible to the radiologists and shouldn't interfere with their workflow. This is important because developing a model alone will not help in integrating them into the clinical application, there must be an interactive medium through which the radiologists can utilise it. Additionally, the model can learn from the interactions of users through the interface, which is another feature that can be beneficial. Using human interactions and feedback in the development of model learning would help to enhance performance and increase the trust of the users in the system.

For both of these goals, we have to build an interactive user-centred platform that can well go with the current medical system they use. The first goal can be obtained by studying the workflow of the radiologists and building a prototype to show them. The second goal is based on the AI and human interactions. The usage of user interactions in retraining also includes collecting data for the same. In other words, while evaluating the reports, we can collect the data in such a way that the annotations can be extracted. This in fact can be turned into an efficient annotation system. Such an annotation process integrated into their workflow also helps us get annotated data efficiently, which is crucial in the medical domain.

Tongshuang et al. [29] conducted a case study tutorial to analyse three aspects of Human AI (NLP) interactions: Usability Evaluation, User interface design and Learning and improving NLP models through human interactions. Learning from human interactions is an interesting aspect that can be adapted to our research. Human-in-the-loop (HITL) in NLP frameworks is an approach in which human feedback is used to improve the model performance which comes under the third aspect of the research mentioned above.

Wang et al. [30] provides a survey on studies that have used HITL in different NLP tasks: Text classification, Parsing and entity linking, Topic Modelling, Summarization and Machine Translation, Dialogue and Question Answering. Among these text classifications and entity linking apply to our research, thus more details on them are discussed further. Models are trained for text classification tasks and users can add or remove text features and label new data if the model performs poorly on them. The same can apply to entity linking, users can interactively annotate entities in text samples.

Lo et al. [31] introduce a similar approach where they use user feedback in the model learning process for entity linking. They used active learning strategies to find the data points that the initially trained model failed to predict and provided them to the users to obtain feedback on whether they were correctly linked or not. This method outperformed the non-interactive baseline models.

To adopt the method of HITL, an interactive medium has to be developed. [32] presents a prototype tool that allows users to visualize and correct the outputs of an NLP system that extracts binary variables from clinical text. There can be two types of interfaces: A graphical user interface where users can interact using windows, buttons, icons, and menus; and a Natural language user interface where they can interact with speech or text like human communication. The feedback that can be collected from the users can also be in various forms:

- Binary user feedback

The feedback will be opposite to each other(Eg: agree, disagree) which can be used in both interactive mediums.

- Scaled user feedback

  The feedback will be in a scaled format like a five-point scale rating.

- Natural language user feedback

  The user can express their feedback in the natural speaking language.

- Counterfactual example feedback

  The feedback is a natural language text like ”If X had not occurred, Y would not have occurred” [30].

Natural language feedback can describe the best intention of the user, but adapting it to the model's understanding is complex when compared to GUI feedback. In addition to the type of feedback, the way of interaction that can be adapted to maximise the performance has also been explored. Active learning mentioned in one of the above studies [31] helps not only for efficient model performance but also reduces the effort of user labelling, by strategically selecting the data that yields the maximum desired output. Another approach is reinforcement learning (RL) where the user interaction is considered an RL action to be taken as a reward or punishment. This supports understanding of human intention while giving feedback by taking them as RL action.

The user feedback can be incorporated into model learning in different ways, such as data augmentation and model direct manipulation. Data augmentation is the process of adding new data samples or features to the data, and the user feedback collected during the interaction can be added as data samples or features. This can happen in two ways, online update and offline update. Offline update of the model retrains the model from scratch after obtaining all the user feedback, while online update trains NLP models while collecting feedback at the same time. Model direct manipulation involves updating the learning parameters, and updating the loss functions of the model through collected numerical user feedback. According to the survey by Wang et al. [30] most of the studies used numerical feedback like binary feedback or scaled feedback for this purpose. However, the inclusion of natural language feedback from users for model direct manipulation can enhance the model learning and performance.

To conclude, previous research has focused on summarizing radiology reports and extracting key descriptors for easier access. However, the application of these methods to ensure the completeness of reports and integrate them into the radiology reporting process has not been thoroughly explored, this is the gap we address in our research. The insights from previous works guided our experimentation with various models and techniques. We explored different NLP models like BERTje, medroBERTa.nl, and CRF, as well as rule-based methods like regular expressions to extract information from the reports. This extraction process was essential in identifying missing information in the reports. We also tried prompting techniques to test whether it is suitable for our task given our limited dataset and also to assess whether existing large language models (LLMs) could be prompted to perform the task effectively. In addition, we explored methods for integrating AI models into workflows and leveraging user interactions for annotation. This is another key focus of this research which demonstrates how keeping the user in the loop can enhance model performance and facilitate the collection of annotated data, which can be applied to other tasks.

Chapter 3

Dataset

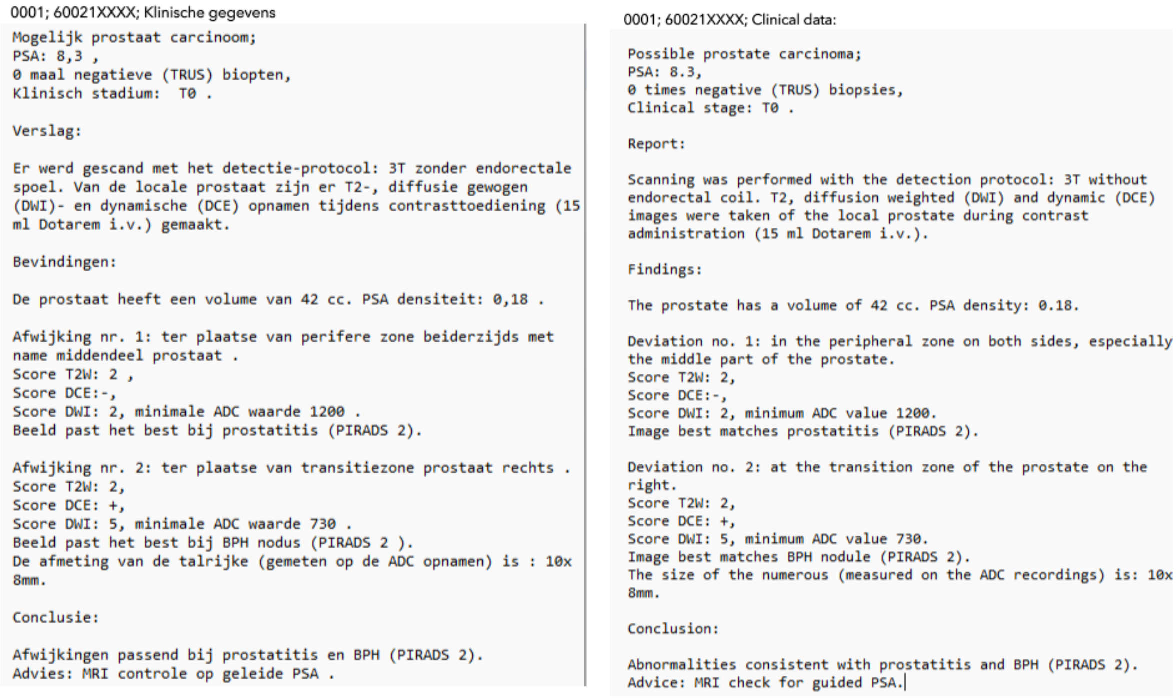
This chapter discusses the dataset we used in this research. This includes the radiology reporting process, how the data is collected and annotated and a detailed analysis of the structure of the report to understand what completeness and compliance refers to.
The chapter is divided into four sections:
Section 3.1: Radiology reporting process
Section 3.2: Dataset Collection
Section 3.3: Dataset Annotation
Section 3.4: Analysis of Report Structure

3.1 Radiology Reporting Process

Radiology reporting is the process of reporting the results of an imaging test. For this research prostate screening reports are taken as the dataset. To understand how they prepared this report, we talked with the radiologists of Ziekenhuisgroep Twente (ZGT). The radiologists analyse the scanning images and report their findings through speech. The hospital has an interface with speech-to-text functionality. The end report will be semi-structured containing the findings and the conclusion. Due to workload and time constraints, the reports can be incomplete and not compliant. Our goal is to automate the check of these reports to ensure their completeness and compliance which saves time and reduces the workload needed for manual checks.

3.2 Dataset Collection

The Dataset used for this research is a set of 206 prostate-scanning radiology reports from Ziekenhuisgroep Twente (ZGT). The reports are semi-structured and contain a section with the patient ID (altered for privacy concerns), a small synopsis of the report, followed by the procedure information, findings and conclusion. Figure 3.1a shows a sample radiology report. The reports are in Dutch and for comprehension purposes Google Translate is used to translate the reports into English which can be seen in figure 3.1b. The reports are from the period March 2015 - January 2017. In addition to the reports, a guideline document was also provided by the hospital: **Quality document Prostate MRI: protocol and reporting** [8]. This guideline document is prepared by the Dutch Association for Radiology (NVvR) and contains the rules of reporting which include the important fields that should be present in a report and how it should be reported. The reports are extracted by the supervisor from the hospital after excluding all the patient info to safeguard their privacy. The reports were used in a secure virtual environment provided by the hospital to ensure data security.



0001; 60021XXXX; Klinische gegevens

```
Mogelijk prostaat carcinoom;
PSA: 8,3 ,
0 maal negatieve (TRUS) biopten,
Klinisch stadium:  T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale
spoel. Van de locale prostaat zijn er T2-, diffusie gewogen
(DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15
ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een volume van 42 cc. PSA densiteit: 0,18 .

Afwijking nr. 1: ter plaatse van perifere zone beiderzijds met
name middendeel prostaat .
Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .
Beeld past het best bij prostatitis (PIRADS 2).

Afwijking nr. 2: ter plaatse van transitiezone prostaat rechts .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .
Beeld past het best bij BPH nodus (PIRADS 2 ).
De afmeting van de talrijke (gemeten op de ADC opnamen) is : 10x
8mm.

Conclusie:

Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .
```

(a) Report

0001; 60021XXXX; Clinical data:

```
Possible prostate carcinoma;
PSA: 8.3,
0 times negative (TRUS) biopsies,
Clinical stage: T0 .

Report:

Scanning was performed with the detection protocol: 3T without
endorectal coil. T2, diffusion weighted (DWI) and dynamic (DCE)
images were taken of the local prostate during contrast
administration (15 ml Dotarem i.v.).

Findings:

The prostate has a volume of 42 cc. PSA density: 0.18.

Deviation no. 1: in the peripheral zone on both sides, especially
the middle part of the prostate.
Score T2W: 2,
Score DCE:-,
Score DWI: 2, minimum ADC value 1200.
Image best matches prostatitis (PIRADS 2).

Deviation no. 2: at the transition zone of the prostate on the
right.
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimum ADC value 730.
Image best matches BPH nodule (PIRADS 2).
The size of the numerous (measured on the ADC recordings) is: 10x
8mm.

Conclusion:

Abnormalities consistent with prostatitis and BPH (PIRADS 2).
Advice: MRI check for guided PSA.
```

(b) Translated Report

Figure 3.1: Sample Prostate Scanning Report

3.3 Dataset Annotation

To train a model for checking the completeness and compliance of reports, we need annotated reports. The annotation in this case is to identify and tag the important fields in the reports. This annotation was carried out by the radiologists of ZGT as they are the experts in identifying the fields in a report. They annotated the reports based on the fields present in the guideline document. The guideline document provides a table of important contents and examples to ensure the good quality of reporting as given in figures 3.2 and 3.3.

The annotated data should be of good quality to obtain a good performance. To ensure this an approach of double annotation was performed. The annotation is done in the inception platform [33] and four radiologists were approached for this task. The reports were divided into two sets (100 and 104), each set of reports was annotated by two of the radiologists each independently resulting in 4 sets of annotated data. After a thorough analysis of the annotated reports, the correct annotations were selected, yielding a total of 204 annotated reports. There were some minor discrepancies between the annotations made by the radiologists, such as missing or swapped annotations. However, these differences were identifiable, and the correct annotations were chosen after consulting with the radiologists for the final dataset.

The radiologists identified 12 fields as important to be presented in the prostate scanning reports. They are given below:

1. PSA waarde (PSA value): Prostate-Specific Antigen (PSA) value obtained from blood tests.

2. Grootte (Size): the size of tumors

3. PSA densiteit MRI (PSA density MRI): Concentration of PSA in the prostate as detected by the MRI.

4. Locatie (Location): Location of specific findings within the prostate.

5. Interpretatie (Interpretation): Interpretation of the imaging findings

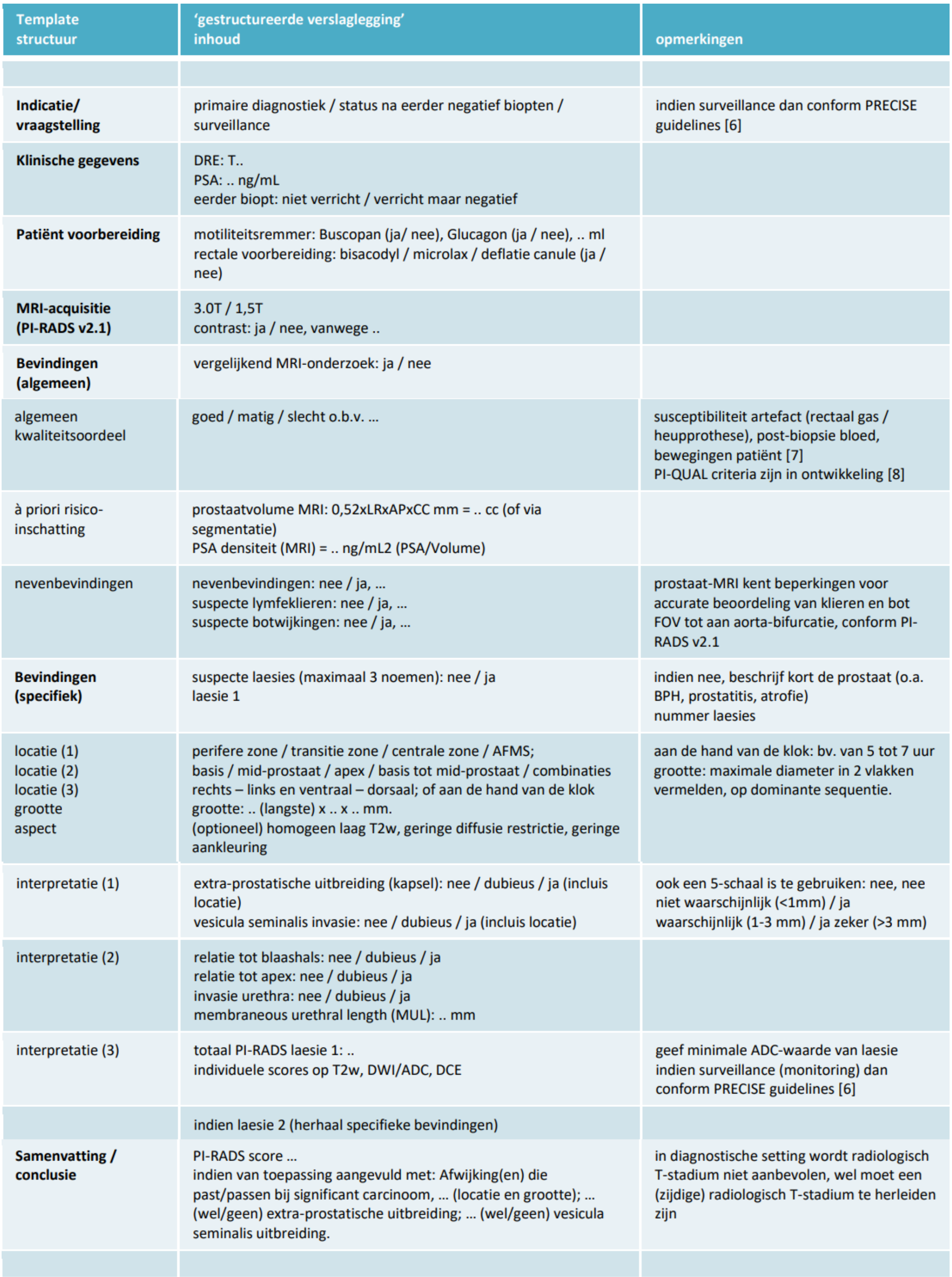
6. Pirads classification: The Prostate Imaging Reporting and Data System (PI-RADS) is a standardized system used to interpret and report findings from prostate MRI. It helps categorize the likelihood of clinically significant prostate cancer based on imaging features.

7. Conclusie (Conclusion)

8. Advies (Advice): Recommendation or advice based on the results of the scan.

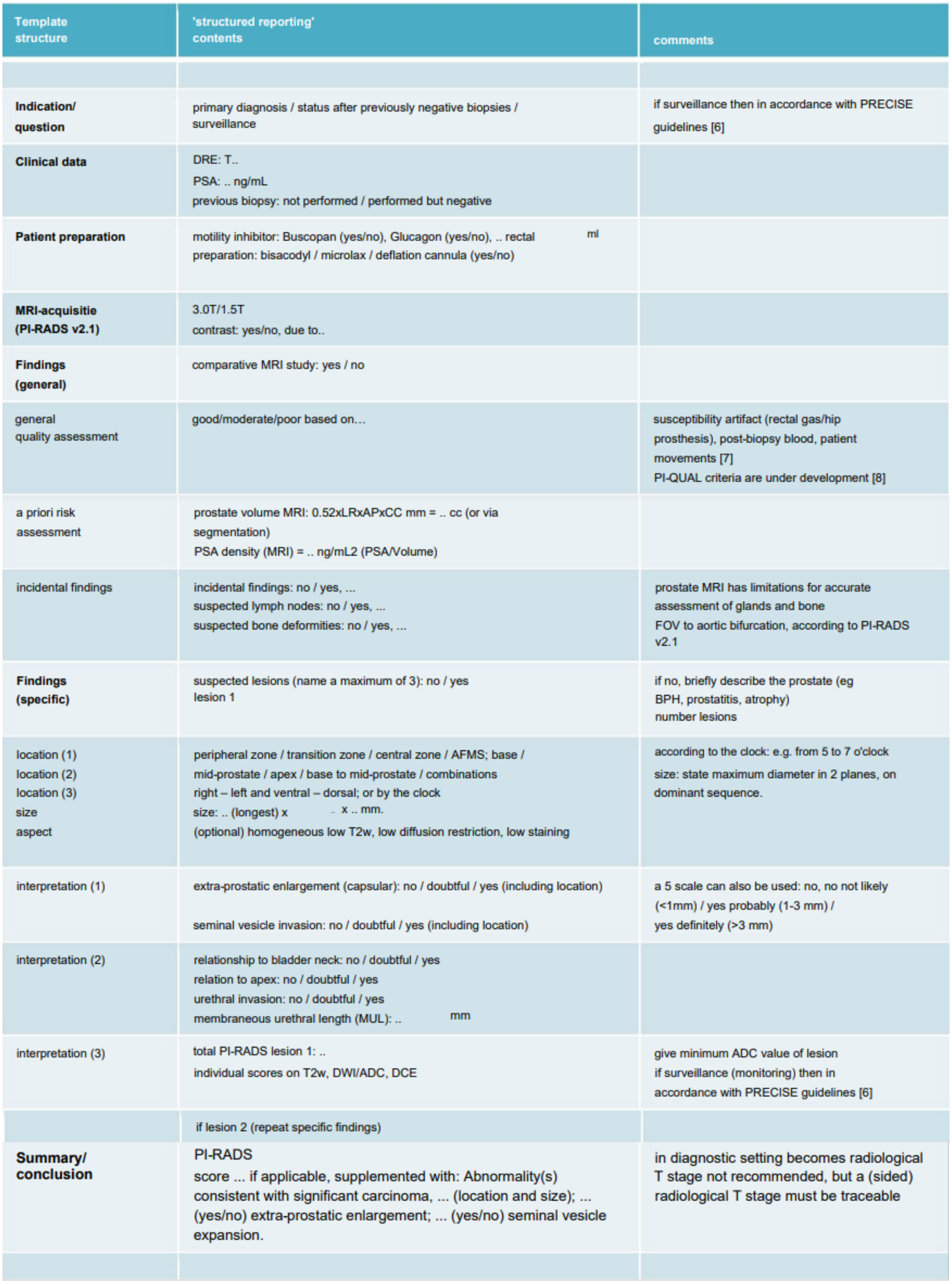
9. Aspect: Appearance or aspect of specific structures or abnormalities within the prostate

10. Kwaliteitsoordeel (Quality rating): Overall quality judgment of the MRI images



Template structuur	'gestructureerde verslaglegging' inhoud	opmerkingen
Indicatie/ vraagstelling	primaire diagnostiek / status na eerder negatief biopten / surveillance	indien surveillance dan conform PRECISE guidelines [6]
Klinische gegevens	DRE: T.. PSA: .. ng/mL eerder biopt: niet verricht / verricht maar negatief	
Patiënt voorbereiding	motiliteitsremmer: Buscopan (ja/ nee), Glucagon (ja / nee), .. ml rectale voorbereiding: bisacodyl / microlax / deflatie canule (ja / nee)	
MRI-acquisitie (PI-RADS v2.1)	3.0T / 1,5T contrast: ja / nee, vanwege ..	
Bevindingen (algemeen)	vergelijkend MRI-onderzoek: ja / nee	
algemeen kwaliteitsoordeel	goed / matig / slecht o.b.v. …	susceptibiliteit artefact (rectaal gas / heupprothese), post-biopsie bloed, bewegingen patiënt [7] PI-QUAL criteria zijn in ontwikkeling [8]
à priori risico-inschatting	prostaatvolume MRI: 0,52xLRxAPxCC mm = .. cc (of via segmentatie) PSA densiteit (MRI) = .. ng/mL2 (PSA/Volume)	
nevenbevindingen	nevenbevindingen: nee / ja, … suspecte lymfeklieren: nee / ja, … suspecte botwijkingen: nee / ja, …	prostaat-MRI kent beperkingen voor accurate beoordeling van klieren en bot FOV tot aan aorta-bifurcatie, conform PI-RADS v2.1
Bevindingen (specifiek)	suspecte laesies (maximaal 3 noemen): nee / ja laesie 1	indien nee, beschrijf kort de prostaat (o.a. BPH, prostatitis, atrofie) nummer laesies
locatie (1) locatie (2) locatie (3) grootte aspect	perifere zone / transitie zone / centrale zone / AFMS; basis / mid-prostaat / apex / basis tot mid-prostaat / combinaties rechts – links en ventraal – dorsaal; of aan de hand van de klok grootte: .. (langste) x .. x .. mm. (optioneel) homogeen laag T2w, geringe diffusie restrictie, geringe aankleuring	aan de hand van de klok: bv. van 5 tot 7 uur grootte: maximale diameter in 2 vlakken vermelden, op dominante sequentie.
interpretatie (1)	extra-prostatische uitbreiding (kapsel): nee / dubieus / ja (incluis locatie) vesicula seminalis invasie: nee / dubieus / ja (incluis locatie)	ook een 5-schaal is te gebruiken: nee, nee niet waarschijnlijk (<1mm) / ja waarschijnlijk (1-3 mm) / ja zeker (>3 mm)
interpretatie (2)	relatie tot blaashals: nee / dubieus / ja relatie tot apex: nee / dubieus / ja invasie urethra: nee / dubieus / ja membraneous urethral length (MUL): .. mm	
interpretatie (3)	totaal PI-RADS laesie 1: .. individuele scores op T2w, DWI/ADC, DCE	geef minimale ADC-waarde van laesie indien surveillance (monitoring) dan conform PRECISE guidelines [6]
	indien laesie 2 (herhaal specifieke bevindingen)	
Samenvatting / conclusie	PI-RADS score … indien van toepassing aangevuld met: Afwijking(en) die past/passen bij significant carcinoom, … (locatie en grootte); … (wel/geen) extra-prostatische uitbreiding; … (wel/geen) vesicula seminalis uitbreiding.	in diagnostische setting wordt radiologisch T-stadium niet aanbevolen, wel moet een (zijdige) radiologisch T-stadium te herleiden zijn

Figure 3.2: Prostate MRI: structured reporting template from the guideline document



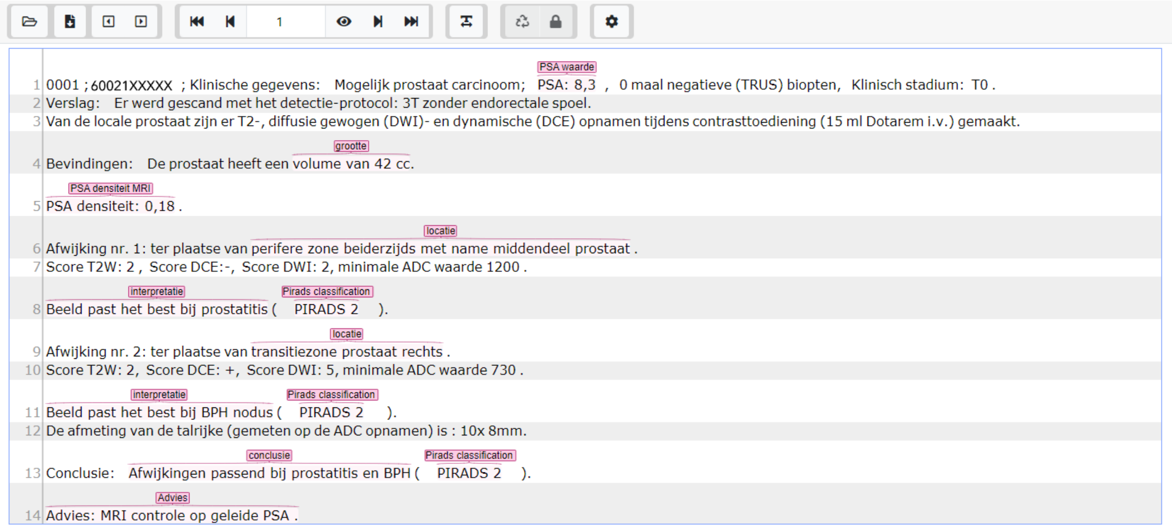
Template structure	'structured reporting' contents	comments
Indication/ question	primary diagnosis / status after previously negative biopsies / surveillance	if surveillance then in accordance with PRECISE guidelines [6]
Clinical data	DRE: T.. PSA: .. ng/mL previous biopsy: not performed / performed but negative	
Patient preparation	motility inhibitor: Buscopan (yes/no), Glucagon (yes/no), .. rectal ml preparation: bisacodyl / microlax / deflation cannula (yes/no)	
MRI-acquisitie (PI-RADS v2.1)	3.0T/1.5T contrast: yes/no, due to..	
Findings (general)	comparative MRI study: yes / no	
general quality assessment	good/moderate/poor based on…	susceptibility artifact (rectal gas/hip prosthesis), post-biopsy blood, patient movements [7] PI-QUAL criteria are under development [8]
a priori risk assessment	prostate volume MRI: 0.52xLRxAPxCC mm = .. cc (or via segmentation) PSA density (MRI) = .. ng/mL2 (PSA/Volume)	
incidental findings	incidental findings: no / yes, ... suspected lymph nodes: no / yes, ... suspected bone deformities: no / yes, ...	prostate MRI has limitations for accurate assessment of glands and bone FOV to aortic bifurcation, according to PI-RADS v2.1
Findings (specific)	suspected lesions (name a maximum of 3): no / yes lesion 1	if no, briefly describe the prostate (eg BPH, prostatitis, atrophy) number lesions
location (1) location (2) location (3) size aspect	peripheral zone / transition zone / central zone / AFMS; base / mid-prostate / apex / base to mid-prostate / combinations right – left and ventral – dorsal; or by the clock size: .. (longest) x .. x .. mm. (optional) homogeneous low T2w, low diffusion restriction, low staining	according to the clock: e.g. from 5 to 7 o'clock size: state maximum diameter in 2 planes, on dominant sequence.
interpretation (1)	extra-prostatic enlargement (capsular): no / doubtful / yes (including location) seminal vesicle invasion: no / doubtful / yes (including location)	a 5 scale can also be used: no, no not likely (<1mm) / yes probably (1-3 mm) / yes definitely (>3 mm)
interpretation (2)	relationship to bladder neck: no / doubtful / yes relation to apex: no / doubtful / yes urethral invasion: no / doubtful / yes membraneous urethral length (MUL): .. mm	
interpretation (3)	total PI-RADS lesion 1: .. individual scores on T2w, DWI/ADC, DCE	give minimum ADC value of lesion if surveillance (monitoring) then in accordance with PRECISE guidelines [6]
	if lesion 2 (repeat specific findings)	
Summary/ conclusion	PI-RADS score ... if applicable, supplemented with: Abnormality(s) consistent with significant carcinoma, ... (location and size); ... (yes/no) extra-prostatic enlargement; ... (yes/no) seminal vesicle expansion.	in diagnostic setting becomes radiological T stage not recommended, but a (sided) radiological T stage must be traceable

Figure 3.3: Translated

11. Prostaatvolume MRI (Prostate volume MRI): Volume or size of the entire prostate as determined by MRI.

12. Vergelijkend MRI onderzoek (Comparative MRI study): Compare this with previous reports.

These fields were used as tags to annotate the reports in the inception platform. An example of an annotated report can be seen in Figure 3.4.



0001 ; 60021XXXXX ; Klinische gegevens: Mogelijk prostaat carcinoom; PSA: 8,3 , 0 maal negatieve (TRUS) biopten, Klinisch stadium: T0 .
Verslag: Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel.
Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.
Bevindingen: De prostaat heeft een volume van 42 cc.
PSA densiteit: 0,18 .
Afwijking nr. 1: ter plaatse van perifere zone beiderzijds met name middendeel prostaat .
Score T2W: 2 , Score DCE:-, Score DWI: 2, minimale ADC waarde 1200 .
Beeld past het best bij prostatitis (PIRADS 2).
Afwijking nr. 2: ter plaatse van transitiezone prostaat rechts .
Score T2W: 2, Score DCE: +, Score DWI: 5, minimale ADC waarde 730 .
Beeld past het best bij BPH nodus (PIRADS 2).
De afmeting van de talrijke (gemeten op de ADC opnamen) is : 10x 8mm.
Conclusie: Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .

Figure 3.4: Annotated Report from the inception platform

The annotated data can be exported from the inception platform as zip files which contain files in XMI (XML Metadata Interchange) format. In these files, annotations are represented by marking the indices of the starting and ending positions of the annotated fields. Figure 3.5 shows an example of the exported annotated data.

```
<type3:NamedEntity xmi:id="2941" sofa="1" begin="74" end="82" value="PSA waarde"/>
<type3:NamedEntity xmi:id="2947" sofa="1" begin="420" end="436" value="grootte"/>
<type3:NamedEntity xmi:id="2953" sofa="1" begin="438" end="457" value="PSA densiteit MRI"/>
<type3:NamedEntity xmi:id="2959" sofa="1" begin="497" end="551" value="locatie"/>
<type3:NamedEntity xmi:id="2965" sofa="1" begin="629" end="664" value="interpretatie"/>
<type3:NamedEntity xmi:id="2971" sofa="1" begin="666" end="674" value="Pirads classification"/>
<type3:NamedEntity xmi:id="2977" sofa="1" begin="714" end="743" value="locatie"/>
<type3:NamedEntity xmi:id="2983" sofa="1" begin="820" end="853" value="interpretatie"/>
<type3:NamedEntity xmi:id="2989" sofa="1" begin="855" end="863" value="Pirads classification"/>
<type3:NamedEntity xmi:id="2995" sofa="1" begin="957" end="999" value="conclusie"/>
<type3:NamedEntity xmi:id="3001" sofa="1" begin="1001" end="1009" value="Pirads classification"/>
<type3:NamedEntity xmi:id="3007" sofa="1" begin="1013" end="1050" value="Advies"/>
```

Figure 3.5: Exported Annotated data

The report with all the tags can be considered complete, if not can be classified as incomplete. This was the initial condition, but on further analysis of the reports, there are some special conditions where the absence of certain tags won't affect the completeness of the reports. This will be discussed in detail in the next section 3.4 The annotation process is carried out as an extra task by the radiologists, and as the radiologists are busy with their work with the hospital, it wasn't ideal to ask them to annotate more amount of reports. Thus the annotated dataset remains of a small quantity.

3.4 Analysis of report structure

This section discusses the detailed analysis of the report structure and contents. This includes the presence of fields, their structure, the frequency of their occurrence and how these points to the completeness and compliance of the reports.

As mentioned above even though the radiologists identified 12 fields for the annotation, there was no report in the given dataset which has all the 12 fields. The highest number of distinct field occurrences in a report is ten and the lowest is four. On further discussion with them, it was found that depending on the findings from the images, the presence of fields can be different. For example *Vergelijkend MRI onderzoek* is a field that will be only present in follow-up reports as it is the comparison of findings with previous reports. The dataset provided were all initial reports, as a result, this field was excluded from the subsequent analysis.

The occurrence of fields is also not unique, that is in a report there are multiple occurrences of the same field. The fields with multiple occurrences in reports are *PSA waarde*, *interpretatie*, *locatie*, *grootte* and *PIRADS classification*. The pattern of their occurrence is also studied to identify them correctly in the case of multiple occurrences. The highest occurring fields are *PIRADS classification*, *PSA waarde*, *interpretatie* and lowest occurring field is *kwaliteitsoordeel*. When asked about the reason for the low occurrence of the *kwaliteitsoordeel*, the radiologist mentioned that often they ignore adding this field if the imaging is of good quality. The frequency of each field can be seen in Table 3.1 and in Figure 3.6.

To understand what contributes towards the completeness and compliance of the reports, the structure and pattern of occurrence of each field are studied.

1. **PSA Waarde (PSA value)**
   This field is reported as PSA: x, where x is the PSA value. Example: PSA: 5

2. **PIRADS classification**
   PIRADS classification is reported as PIRADS x, where x is the classification which is written as both in numerical digits as well as Roman numbers. Example: PIRADS 5, PIRADS V. This field also occurs multiple times, and each time can be the same or different. PIRADS value changes according to the lesions present in the prostate. So if there is more than one lesion, then there

Table 3.1: Field Frequency table

	Field	**Frequency (Number of reports with this field)**
1	PIRADS classification	204
2	PSA waarde	204
3	Interpretation	204
4	Conclusie	202
5	Prostaatvolume MRI	200
6	PSA densiteit MRI	200
7	Locatie	200
8	Advies	127
9	Grootte	67
10.	Kwaliteitsoordeel	4
11	aspect	8
12	Vergelijkend MRI onderzoek	0

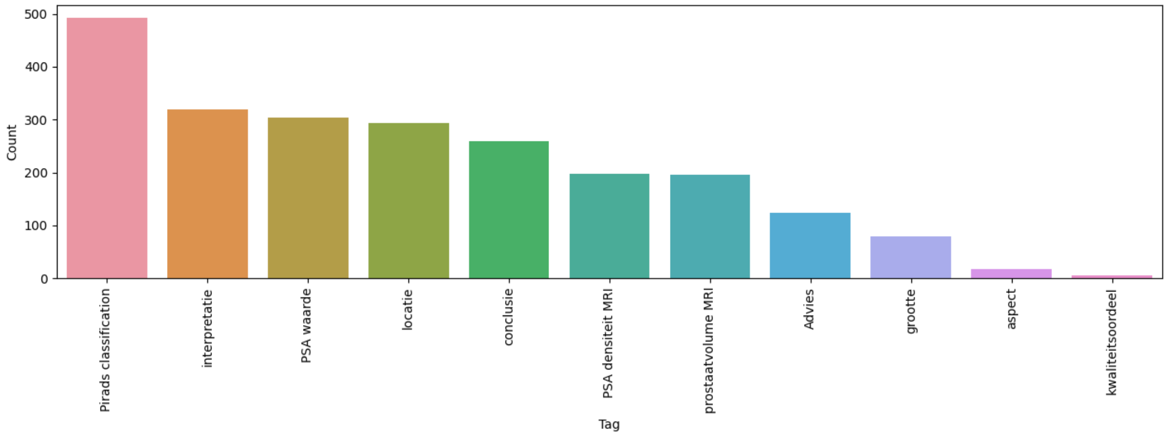


Figure 3.6: Frequency of Tags

will be more than one abnormality (afwijking) and each can result in a different PIRADS classification.

3. **PSA densiteit MRI (PSA density MRI)**
   This field is reported as PSA densiteit: x, where x is the value. According to the guideline, there is also a metric 'ng/mL2' present along with the value, but radiologists don't use that because it is already understood without mentioning it. This field only occurs once. Example: PSA densiteit: 0.24.

4. **Prostaatvolume MRI (Prostate volume MRI)**
   This field is reported as Prostaatvolume: x cc, where x is the value and cc is the standard metric. This field only occurs once in a report. Example: Prostaatvolume 22 cc.

5. **Interpretatie (Interpretation)**
   This field doesn't have a defined structure. It provides the interpretation of the findings and also occurs more than once based on the number of abnormalities and lesions. The interpretation also contains the Pirads classification sometimes.

   Example: Beeld past het best bij prostatitis (PIRADS 2).

   Translation: Image best matches prostatitis (PIRADS 2)

6. **Locatie (Location)**
   This field also doesn't have a defined structure, they contain the zones where the lesions are present. It often occurs along with the interpretation field and occurs more than once based on the abnormalities and lesions.

   Example: ter plaatse van perifere zone beiderzijds met name middendeel prostaat.

   Translation: at the peripheral zone on both sides, especially the middle part of the prostate.

7. **Grootte (Size)**
   This field is reported as a x b mm, where a and b are the size values and mm is the standard metric. This can also occur more than once based on the number of abnormalities and lesions. Example: 10x8 mm.

8. **Aspect**
   This field doesn't have a defined structure and is only present in some of the reports because it represents the structural abnormalities or aspects of specific structures, which will not be always present in a report.

   Example: Geen betrokkenheid van de zaadblaasjes.

Translation: No involvement of the seminal vesicles.

9. **Conclusie (Conclusion)**
   This also doesn't have a defined structure, but it starts with a conclusie and a summary of the report. According to the guidelines it should also contain the PIRADS value if applicable. However, there can be different PIRADS classifications in a single report. On further discussion with the radiologist, it was decided that the highest PIRADS value is the one that should be included in the conclusion.

   Example: Conclusie: Afwijkingen passend bij prostatitis en BPH (PIRADS 2).

   Translation: Conclusion: Abnormalities consistent with prostatitis and BPH (PIRADS 2).

10. **Advies (Advice)**
    This field also doesn't have a structure except it often starts with Advies. This also doesn't occur in every report, only in those which need advice.

    Example: Advies: MRI controle op geleide PSA .

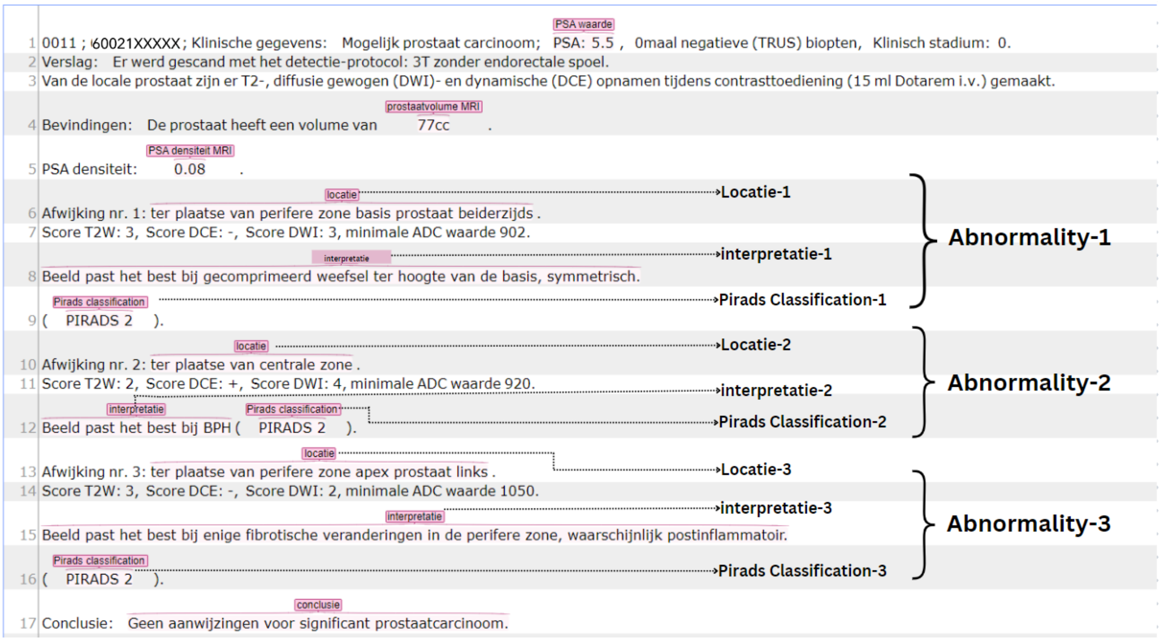
    Translation: Advice: MRI check for guided PSA.

11. **Kwaliteitsoordeel (Quality rating)**
    This field also doesn't have a structure and has the lowest frequency. As mentioned before the radiologist only reports the quality when it is bad. Describe the quality of MRI which contains words like susceptibiliteit artefact, Matige kwaliteit scan (poor quality scan).

After the analysis of the dataset, it is understood that the presence of all fields is not necessary, some fields are more important than others as indicated by their frequency of occurrence. This was confirmed by the radiologists, as they mentioned that some fields are not always present in the reports like *advies*, *aspect*. Also, the presence of certain fields depends on the report findings, as the absence of a tumour will result in the absence of fields like *locatie* and *grootte*.

Also, there are multiple occurrences of the same field in each report, emphasizing the need to study the pattern of these occurrences. The fields' structure can help identify and create the features necessary for developing the model. Now that we have analysed the field structure we should also see how the report is structured.

The radiologist report their findings in an order and that itself is the structure of the report. In figure 3.7 you can see how the report looks like when there are multiple occurrences of fields. After observing the reports, we were able to identify that the three fields - *interpretatie*, *locatie*, and *PIRADS classification* occur on each



```
0011 ; 60021XXXXX; Klinische gegevens:   Mogelijk prostaat carcinoom;  PSA: 5.5 ,  0maal negatieve (TRUS) biopten,  Klinisch stadium:  0.
Verslag:   Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel.
Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.
Bevindingen:   De prostaat heeft een volume van      77cc      .
PSA densiteit:      0.08      .
Afwijking nr. 1: ter plaatse van perifere zone basis prostaat beiderzijds .
Score T2W: 3,  Score DCE: -,  Score DWI: 3, minimale ADC waarde 902.
Beeld past het best bij gecomprimeerd weefsel ter hoogte van de basis, symmetrisch.
(   PIRADS 2    ).
Afwijking nr. 2: ter plaatse van centrale zone .
Score T2W: 2,  Score DCE: +,  Score DWI: 4, minimale ADC waarde 920.
Beeld past het best bij BPH (   PIRADS 2    ).
Afwijking nr. 3: ter plaatse van perifere zone apex prostaat links .
Score T2W: 3,  Score DCE: -,  Score DWI: 2, minimale ADC waarde 1050.
Beeld past het best bij enige fibrotische veranderingen in de perifere zone, waarschijnlijk postinflammatoir.
(   PIRADS 2    ).
Conclusie:   Geen aanwijzingen voor significant prostaatcarcinoom.
```

Figure 3.7: Report structure showing Multiple occurrences of fields

identification of abnormality or lesion. So that is a pattern for their multiple occurrence. Also, the order in which the fields occur is almost the same in most of the reports. *PSA waarde*, *prostaatvolume MRI*, *PSA densiteit MRI* on top followed by the *locatie*, *interpretatie* and *PIRADS classification* in the findings. Then the *grootte*, *aspect* and *advies* are present if necessary according to the findings and *conclusie* at the end.

What makes the report complete and compliant? This is the question that should be addressed before delving into the models to ensure them. The initial standard we came up with to describe completeness and compliance was, that a report is considered to be complete if it has all the fields present and it is considered compliant if the field contents follow the structure according to the guideline (Refer to column ”structured reporting contents” in figure3.3). But on further analysis of the dataset and discussion with the radiologists, we decided that completeness cannot be ensured by the presence of all the fields, because the report doesn’t need to contain all the fields, it changes based on the findings. So to make sure the reports are complete, we can find the missing fields and suggest them to the radiologists, and they can add them if felt necessary. That is the missing fields will be suggestions that the radiologists can either accept or ignore. So to check if a report is complete we identify the fields and present the missing ones to the radiologists, they are the ones who decide if they want to add the fields to the report and make it complete.

When coming to compliance, the guideline rules were taken into consideration, but later on, the rules were found not applicable in certain cases, which is discussed in detail in the next sections 4.1.4.

Chapter 4

Identifying the Missing Fields and Ensuring Compliance of Reports

Following the dataset analysis, the next step is to develop methodologies to address our research questions. This chapter describes our journey to develop an approach for predicting missing fields to ensure the completeness of radiology reports, followed by defining rules to ensure compliance. This constitutes our solution to the first research question. This chapter outlines the methodology used, the experiments conducted, and the results obtained.

4.1 Methodology

This section describes the methodology we adopted to identify the missing fields and ensure the reports' compliance. It provides an overview of the models and techniques we utilised.

To identify the missing fields first, we need to detect the fields present in the reports. Given that the annotation process tags these fields, we can train models to recognize the important fields. In the context of NLP, Named Entity Recognition (NER) is an appropriate task for this purpose. The reports are tokenized and each token is labelled with the corresponding entity. Based on a thorough review of related studies, we experimented with three different methodologies to perform NER and identify the fields in the reports. Once the important fields are identified, missing fields can be determined by noting which expected fields are absent from the identified ones.

The following three subsections describe the different approaches we tested to ensure the completeness of reports:

4.1.1: BERT and RoBERTa-based Dutch Language Models

4.1.2: Prompting

4.1.3: Hybrid Conditional Random Field

After identifying the missing fields the compliance of the reports which are based on the guideline rules will be discussed in the subsection 4.1.4.

Data Pre-processing

Before delving into the models - some common pre-processing steps were taken to ensure the smooth functioning of the models. The annotated versions of the reports were extracted as XML files. These annotations are represented by marking the indices of the starting and ending positions of the tags (fields) (see figure 3.5). For training the models these annotations were converted into BIO (Beginning, Inside, Outside) format. These BIO-formatted tokens are used for the training of the models. These preprocessing steps were used in the following approaches.

4.1.1 Language models

Previous research has demonstrated that leveraging Language Models (LMs) can effectively identify important parts of a report. Given that our reports are in Dutch, utilizing Dutch language models is a logical choice. So the following LMs were used to identify the important fields by using the training data. Due to the limited annotated dataset, we expected the challenge of overfitting, so prompting the models was also considered. We proceeded by selecting models trained on Dutch medical notes and data.

We chose MedRoBERTa.nl, a medical domain-specific LM which has demonstrated good performance on similar tasks before. We also experimented with the Dutch LM BERTje which is pre-trained on general data as one of the previous studies showed that the general LMs can sometimes outperform domain-specific ones. However, despite our efforts to mitigate it, overfitting remained an issue with these models. Figure 4.1 shows the general flow of the training process in the LMs.

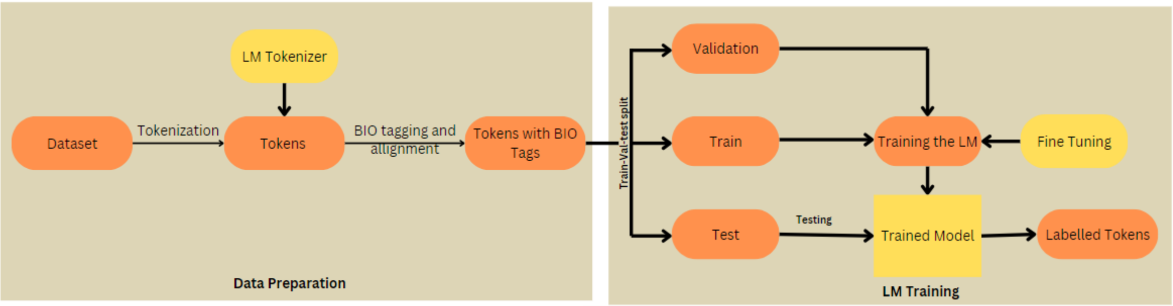


Figure 4.1: Flow chart

The following sections provide a discussion of these language models and their respective training processes.

BERTje - Dutch BERT Model

BERTje is a Dutch pre-trained BERT developed at the University of Groningen [34]. BERTje is pre-trained on various corpora of high-quality Dutch text and can be used for various NLP tasks like next-sentence prediction, POS tagging, and NER.

NER is the task we have to perform. Bidirectional Encoder Representations from Transformers (BERT) based models can capture the bidirectional contexts by considering both the left and right contexts of each word. In our case, this will be helpful, especially in fields with more than two words like interpretation, location, aspect, advice and conclusion. A previous study by Rietberg et al. [35] also shows that BERTje outperformed the domain-specific language model MedRoBERTa on the task of extracting the reason for taking an MRI scan of Multiple Sclerosis (MS) patients, proving domain-specific models are not always the superior ones. Thus we tried the BERTje model on our task to see the possibility of it performing well.

Tokenization

The training procedure started with the general preprocessing, then the BERT tokenizer was used to tokenise the data. We used the word tokenized reports as the data instead of the raw report text to keep track of the labels. But the problem was the tokenizer vocabulary didn't have most of the words for our domain, and the BERT tokenizer treated them as subwords, which is one of the features of BERT models. Initially, the thought of adding the domain-specific tokens into the vocabulary seemed like a good approach. However, it turned out that the number of these specific words was relatively small compared to the overall vocabulary of the tokenizer. As a result, this addition did not significantly impact the model's performance. So the original BERT tokenizer was used and the label alignment was done after the tokenization.

Detailed analysis of the training parameters and results of the model performance can be seen in the coming section 4.2.1.

MedRoBERTa.nl- Dutch Medical LM

MedRoBERTa.nl is the first Dutch language model for the medical domain [20]. Since we have a Dutch medical dataset, this model was selected for our task. MedRoBERTa.nl is trained from scratch with Dutch hospital notes and has outperformed general language models for Dutch in classifying sentences in Dutch hospital notes. This model is based on the RoBERTa architecture, which is a transformer model and is known for its effectiveness in NLP tasks.

Tokenization

The tokenizer used here is ”CLTL/MedRoBERTa.nl” but unlike the BERTje tokenizer, this tokenizer has an expanded vocabulary specifically tailored to the medical domain, resulting in more effective and accurate tokenization of medical terms and phrases. The training parameters and result analysis are discussed in section 4.2.1

4.1.2 Prompting

Prompting is an approach where models are prompted with defined questions or prompts tuned to get particular results from the model. We opted for this approach because of the limited annotated data available. But the LMs like BERT and RoBERTa-based model prompting without pretraining isn't a good approach as the models do not have any idea of our particular domain which is prostate screening reports.

The other option was to try prompting on large language models like GPT, and Mistral AI. However, using publicly available LLMs is not an ideal approach because of the sensitive data we have. Thus to understand the upper bound, we decided to prompt Mistral AI to see how it can perform the task. The Mistral AI 7B model performed well in areas like mathematics, code generation, and reasoning [36]. Thus this LLM was chosen for prompting. Since the annotated data is anonymised, using this for prompting was not an issue, we also got permission from the hospital supervisor.

Two types of prompting: zero-shot and few-shot prompting were performed on Mistral AI.

1. **Zero-shot**

   For the zero-shot prompting, the following prompt was used:

   Prompt : *Identify the following fields from the dutch prostate screening report based on the given guideline: conclusie, Advies,locatie, grootte, aspect, prostaatvolume MRI ,interpretatie', Pirads classification, kwaliteitsoordeel, PSA densiteit MRI, PSA waarde.*

   The report and the guidelines table were given along with the prompt.

2. **Few-shot**

   For the few-shot prompting, an example of how the result should look was also given.

   Prompt: *Identify the following fields from the dutch prostate screening report based on the given guideline: conclusie, Advies,locatie, grootte, aspect,* *prostaatvolume MRI ,interpretatie', Pirads classification, kwaliteitsoordeel, PSA densiteit MRI, PSA waarde. Here is an example The report : The guideline: The labels*

The experiments and results are discussed in the following section 4.2.2

4.1.3 Hybrid Conditional Random Field

The Conditional Random Field (CRF) is a probabilistic model that is used for sequence labelling tasks. The Named Entity Recognition (NER) required to identify the fields in the reports can be performed by CRF. Previous research has used CRF to extract information from health records and medical notes. We also require the same functionality here to label the fields in the report.

The Conditional Random Field is a discriminant model that models the conditional probability of the output labels given the input features $P(Y|X)$, where X represents the input features in our case the tokens and the features related to it, and Y represents the output labels which will be the field tags [37].

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_i, y_{i-1}, X, i)\right)$$

where,

f_k are feature functions
λ_k weights for the feature functions
$Z(X)$ is the normalization factor to ensure that the probabilities sum to 1

CRFs model the dependencies between labels in sequence considering both the current and neighbouring labels, which helps in making accurate predictions in sequences. Before training the model, we can manually add the necessary features to the CRF model. This can be done by carefully analysing the field structures and recognising their pattern. However, the addition of features had a negative impact on the performance. The reason may be due to the fact the new features have reduced the generalizing ability of the model. So just the general features regarding the font case and title were added.

After incorporating the features, the data is fed into the model for training. The trained model is then utilized to predict the tags for the test data. The training process involves using 80% of the dataset, with the remaining 20% reserved for testing. The parameters of the CRF model were experimented with, and the final parameters that gave the best performance were selected.

Since regular expression rules can easily define some of the patterns, they were added as a post-processing step, to make sure they are captured even if the model fails to. Table 4.1 shows the regular expression rules formulated to improve the labelling. Other than these rules, some extra rules were given to identify the multiple occurrences of fields Interpretatie, Locatie, and Pirads classification. The occurrence dependency of these fields is based on the abnormality count that has been described in the dataset section (see figure 3.7). So rules were made based on the number of abnormalities or lesions (Afwijking).

Post-processing of the label was taken into account for other reasons too. After the evaluation with a new set of evaluation data from the hospital, it was noticed that there were some changes in the way some of the fields were reported. For example, the PSA densiteit MRI field was reported in the given training dataset as PSA densiteit: 0.7, but in the new evaluation data it is also reported as density: 0.7. These small changes are not identified by the model, because of their absence in the training data. So defining some rules to incorporate such changes by identifying the pattern can increase the performance.

Thus our final model is a hybrid CRF model combining the power of both the CRF and the regular expression rules. Figure 4.2 shows the architecture of the Hybrid CRF model we used.

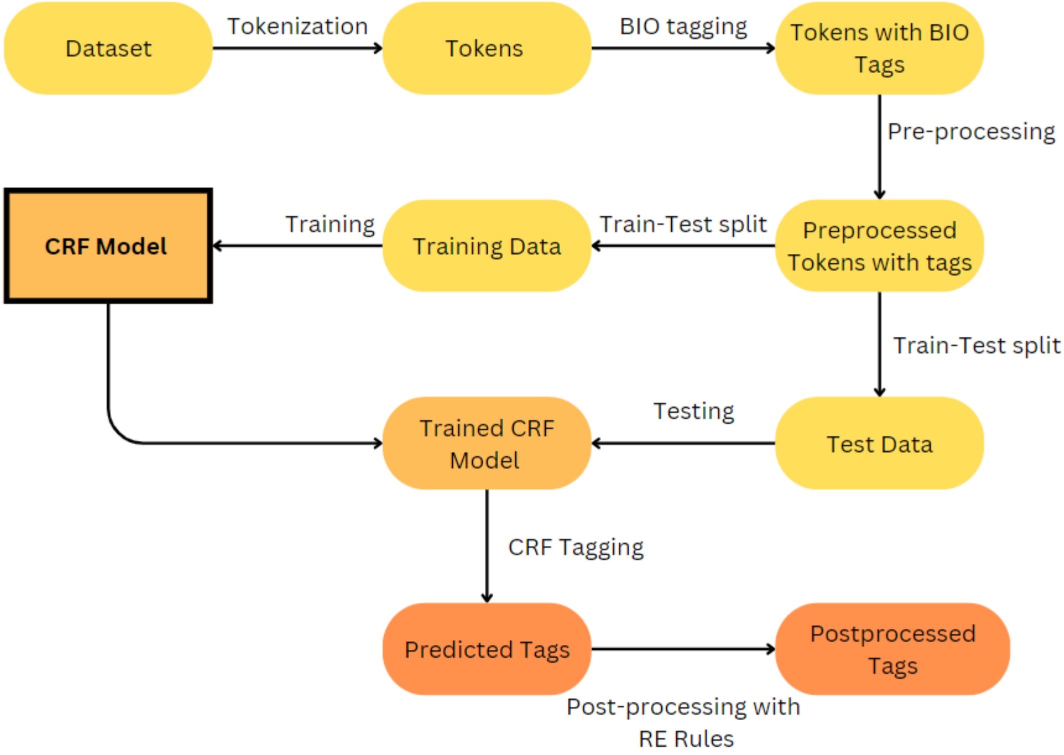


Figure 4.2: Hybrid CRF model

After identifying the fields in the reports, the missing fields can be determined by

Field	**Condition**	**Regex**
Conclusie (Conclusion)	Detect the word *conclusie:* or *conclusie*	`token.lower() == ’conclusie’` or `token.lower() == ’conclusie:’`
PSA Waarde(PSA Value)	Detect PSA value in the format *PSA: x*	`re.match(r’PSA:\d+’, token)`
PSA densiteit MRI	Detect *PSA density:x* or *PSA denistiet*	`token.lower() in [’densiteit’, ’densiteit:’, ’density’, ’density:’]` followed by a decimal check with is_decimal() function

Prostaat volume MRI	Detect *volume: x cc or volume: x ml*	`re.match(r’\^ \d+([.,]\d+)?\s?cc$’, token)` or `re.match(r’\^ \d+([.,]\d+)?\s?ml$’, token)`
		`token.lower() == ’volume’` followed by a decimal check and `token[i+2] in [’cc’, ’ml’]`

Advies(Advice)	Detect the word *Advies:* or *Advies*	`token.lower() == ’advies’` or `token.lower() == ’advies:’`
PIRADS Classification	Detect *pirads* followed by a digit	`token.lower() == ’pirads’` and `tokens[i+1].isdigit()`
Grootte (Size)	Detect size in *mm* or *cm* format	`re.match(r"\b \d+mm \b", token)` **or** `is_decimal(token)` followed by `tokens[i+1] in [’mm’, ’cm’]`

Table 4.1: Post-Processing Rules for Label Adjustment

searching for those not present in the identified fields set.

4.1.4 Compliance

After identifying the missing fields our next task is to look for compliance issues in the report. At the start, the report's compliance was decided based on the table from the guideline document 3.3. However, after further review of the reports, it was noticed that the field content doesn't exactly have the structure provided in the guideline table. The radiologist mentioned that in the guidelines the examples of the fields, are given rather than the structure of it. So to decide on the rules that are required for the compliance of the report, we discussed with the radiologists and fixed what rules are necessarily required to make the report compliant. Only some of the fields had the rules to be followed.

To describe what we meant by compliance in this research is that after the report is complete with the required fields, the fields present in the report should abide by certain criteria. So to ensure the compliance of the report we set some rules for certain fields.

1. Prostaat Volume MRI - The value and the metric cc
2. PSA densiteit MRI - The value and the metric ng/mL2
3. Conclusie - The conclusion should have the pirads value (the one with the highest score) if applicable.

The other fields do not have specific rules to adhere to and therefore only these three fields were considered while making the rules. The given reports follow these rules, except for the PSA densities field where the metrics are not included. Also during the evaluation, the evaluation dataset showed some variation in reporting prostrate volume where ml was used as the metric instead of cc.

Compliance with the report doesn't have many strict rules to be followed, only the three mentioned, but the two of them are often not used, because of the routine, the users of the radiology reports know what they are dealing with. However we decided to include these rules and provide them as suggestions to the user to ensure the reporting goes compliant as mentioned in the guideline document.

4.2 Experiment and Results

This section discusses the experiments conducted with the models and presents their results and analysis. This also addresses the challenges we faced and finally, highlights the methodology that proved most effective for our problem.

4.2.1 Language Models– Results

BERTje

Training the Model– Hyperparmeter tunings

The tokenized dataset was divided into 3 sets: Train, Validation and Test dataset in the proportion 80%-10%-10%. After experimenting with various configurations, the following hyperparameters were identified as the best for training the BERTje model.

- Number of Epochs: 30
- Batch Size: 8, this specifies the number of samples processed before the model's internal parameters are updated.
- Learning rate: initial value 5e-5
- Optimizer: AdamW (Adam with Weight Decay) is used as the optimizer
- Scheduler: A linear scheduler (lr) with warmup is used, thus the lr is linearly decreased after the warmup.

The 30-epoch training resulted in an overfitting with test loss increasing after a point. To address overfitting, we replaced the fixed learning rate (LR) parameter with a learning rate scheduler to enhance the model's ability to generalize. Additionally, we implemented early stopping to halt the training process as soon as overfitting was detected by keeping track of validation loss. This was configured to monitor the validation performance, stopping the training process when there was no further improvement in the validation set.

After these changes, despite the effort, the model continued to exhibit overfitting. The early stopping halted the training by the 11th epoch, indicating the onset of overfitting. Figure 4.3 shows the trends in training and validation loss, as well as the training and validation accuracy over the epochs. The accuracy is measured on the token level of the fields which reflects the proportion of tokens that were correctly classified out of all the tokens. Before overfitting occurred, the model achieved a training accuracy of 76% and a validation accuracy of 71%.

MedRoBERTa.nl

Training the Model - Hyperparmeter tunings

The tokenized dataset was divided into 3 sets- Train- Validation and Test dataset in the proportion 80%-10%-10%. The following hyperparameters were utilised after experimenting with different configurations for the training of the MedRoBERTa.nl model.

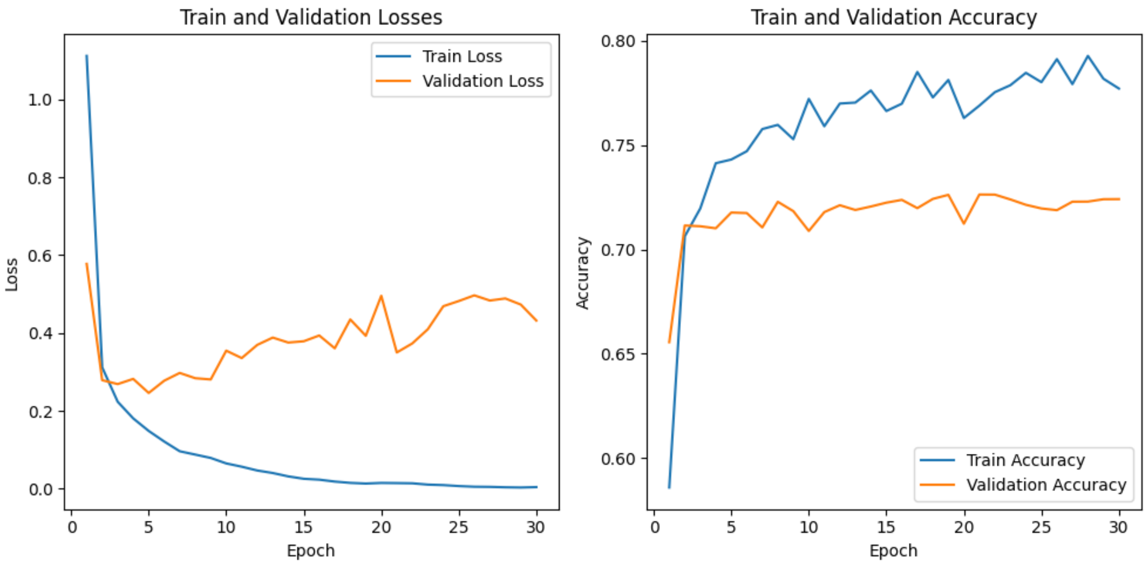


Figure 4.3: BERTje: Train- Validation Loss and Accuracy vs Epochs

- Number of Epochs:30
- Learning rate: initial 5e-5
- Batch size: 4
- Optimizer: AdamW
- Scheduler: ReduceLROnPlateau with Mode min that will reduce the lr when validation loss stops decreasing and patience of 2, the number of epochs with no improvement after which lr will be reduced.

This combination of hyperparameters was identified as the most effective through multiple trials. The model is trained for up to 30 epochs. Attempts were made to address the overfitting issue by adjusting the learning parameters. This included experimenting with various learning rate schedules to improve model generalization. However, these adjustments did not lead to significant improvements or reduce the overfitting problem.

Figure 4.4 shows the train Loss and validation loss, along with the train and validation accuracy. Before overfitting the train accuracy is 79% and validation accuracy is 76%. This model performed better than the BERTje. This suggests that with effective strategies to mitigate overfitting, MedRoBERTa.nl could potentially deliver even better performance.

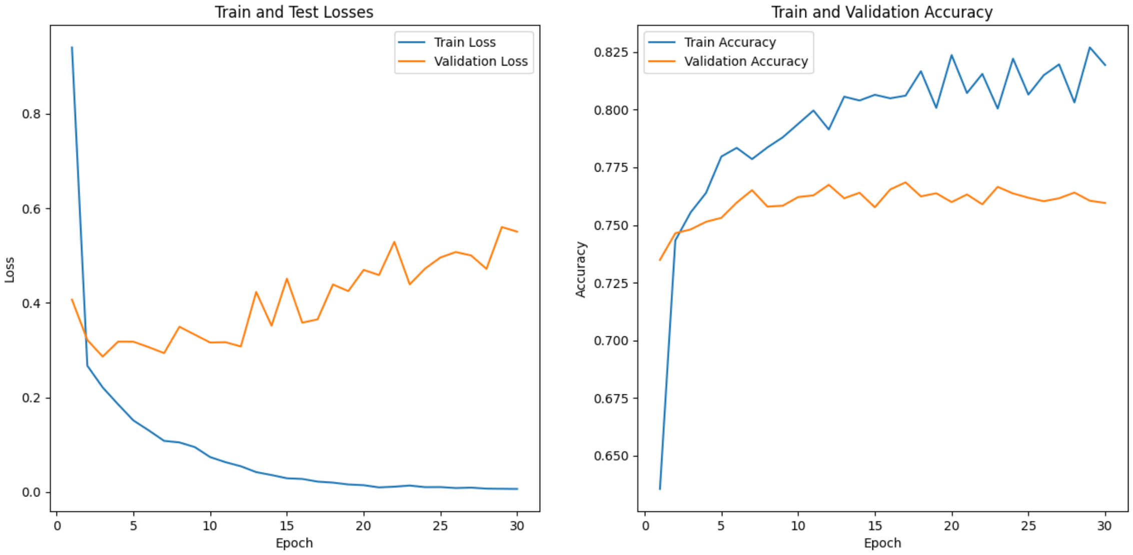


Figure 4.4: MedRoBERTa.nl: Train- Validation Loss and Accuracy vs Epochs

4.2.2 Prompting

Prompting with Mistral AI gave good results. The few-shot approach was more suitable as the structure of the output was defined and easier to compare with the true labels we have. Consequently, we applied this approach across 204 reports.

We compared the results against the annotated reports through qualitative analysis. The quantitative analysis did not provide a strong evaluation of the model's output, as there were slight discrepancies between the identified fields and the annotated data. For example, while the annotated data labelled “PSA densiteit MRI” as “PSA densiteit: 0.25,” Mistral AI identified the field as “0.25”. This is technically correct but differs in format. Similar variations were observed in other fields, thus we manually reviewed 40 reports to assess whether Mistral AI successfully identified all relevant fields.

The following shows an example output of Mistral AI that is correctly identified :

Mistral AI output

Report ID: 032

{’Value’: ’PSA waarde’, ’CoveredText’: ’PSA: 7.6’}

{’Value’: ’prostaatvolume MRI’, ’CoveredText’: ’60 cc’}

{’Value’: ’PSA densiteit MRI’, ’CoveredText’: ’0,13’}

{’Value’: ’locatie’, ’CoveredText’: ’ter plaatse van perifere zone posterolateraal mid en apex prostaat links’}

{’Value’: ’interpretatie’, ’CoveredText’: ’Beeld past het best bij significant carcinoom (PIRADS 5)’}

{’Value’: ’Pirads classification’, ’CoveredText’: ’PIRADS 5’}

{’Value’: ’grootte’, ’CoveredText’: ’11 x 16 mm’}

{’Value’: ’conclusie’, ’CoveredText’: ’Afwijking passend bij significant carcinoom links perifere zone posterolateraal mid en apex prostaat (PIRADS 5)’}

{’Value’: ’Advies’, ’CoveredText’: ’MRI geleid biopt’}

Annotation (Ground truth)

Report ID: 032

{’Value’: ’PSA waarde’, ’CoveredText’: ’7,6’}

{’Value’: ’prostaatvolume MRI’, ’CoveredText’: ’60 cc’}

{’Value’: ’PSA densiteit MRI’, ’CoveredText’: ’0,13’}

{’Value’: ’locatie’, ’CoveredText’: ’ter plaatse van perifere zone posterolateraal mid en apex prostaat links’}

{’Value’: ’interpretatie’, ’CoveredText’: ’Beeld past het best bij significant carcinoom’}

{’Value’: ’Pirads classification’, ’CoveredText’: ’PIRADS 5’}

{’Value’: ’grootte’, ’CoveredText’: ’11 x 16’}

{’Value’: ’conclusie’, ’CoveredText’: ’Afwijking passend bij significant carcinoom links perifere zone posterolateraal mid en apex prostaat’}

{’Value’: ’Pirads classification’, ’CoveredText’: ’(PIRADS 5’}

{’Value’: ’Advies’, ’CoveredText’: ’MRI geleid biopt’}

However, there were cases where Mistral AI incorrectly identified certain fields. For example, a random text in the report was identified as advies by Mistral AI, where the report doesn’t have the field advies.

Report ID: 028

{’Value’: ’Advies’, ’CoveredText’: ’Patiënt werd ingepland voor MRI geleid biopt in het kader van de studie.’}

Another incorrect example occurred with the ”kwaliteitsoordeel” field, where a random piece of text was mistakenly labelled as this field, even though the report did not contain ”kwaliteitsoordeel”:

Report ID: 014

{’Value’: ’kwaliteitsoordeel’, ’CoveredText’: ’Het onderzoek werd medebeoordeeld door UMCN prof. Barentsz/ M vd Leest’}

The discrepancies in the true and predicted labels when analysed showed that the Mistral AI has fabricated some of the fields that were not present in the report. These can be hallucinations or the model has assigned some other texts in the report as the field. This especially occurred with the fields advies and kwaliteitsoordeel as seen in the above examples. Additionally, in our manual analysis of the 40 reports, we found cases where Mistral AI failed to identify fields that were present, especially ”kwaliteitsoordeel” and ”aspect.” To conclude while Mistral AI was able to correctly identify some fields, it also fabricated certain fields and missed others. These findings, along with the privacy concerns associated with using public AI models, suggest that such tools may not yet be suitable for use in the medical domain.

4.2.3 Hybrid Conditional Random Field

To evaluate the identification of the fields using the CRF model, two approaches were taken

1. Quantitative Analysis

   For the Quantitative analysis, the following evaluation metrics were used: Precision gives the number of true positives given all the positive cases, whereas recall gives the number of positives given all the true positives and false negatives. F1-score combines precision and recall and is useful in labelling tasks.

   $$Precison = \frac{TP}{(TP + FP)}$$

   $$Recall = \frac{TP}{(TP + FN)}$$

   $$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

   where,

   TP: True Positive

   FP: False Positive

   FN: False Negative

   We performed two types of token-level evaluations:

   - Detailed Evaluation: Assesses each B-token (the beginning token of an entity) and I-token (the inside token of an entity) labels individually. Each token in the field is evaluated separately.
   - Combined Evaluation: Evaluate the labels considering both B and I tokens together that give the evaluation of the field as the whole text.

   Another evaluation metric is the metrics seqeval. But it is a stricter metric, and might not fully capture the model's performance for our task, because it evaluates the entire sequence at once. As a result, even slight changes or mistakes in the text can lead to a larger penalty, potentially misrepresenting the actual performance of the model. Hence, we focused on the detailed and combined evaluations. Figure 4.5 shows the detailed evaluation metrics of each label and figure 4.6 shows the combined evaluation metrics of each label. To have a more clear understanding the F1-score is separately mapped for both cases and is shown in figures 4.7 and 4.8respectively.

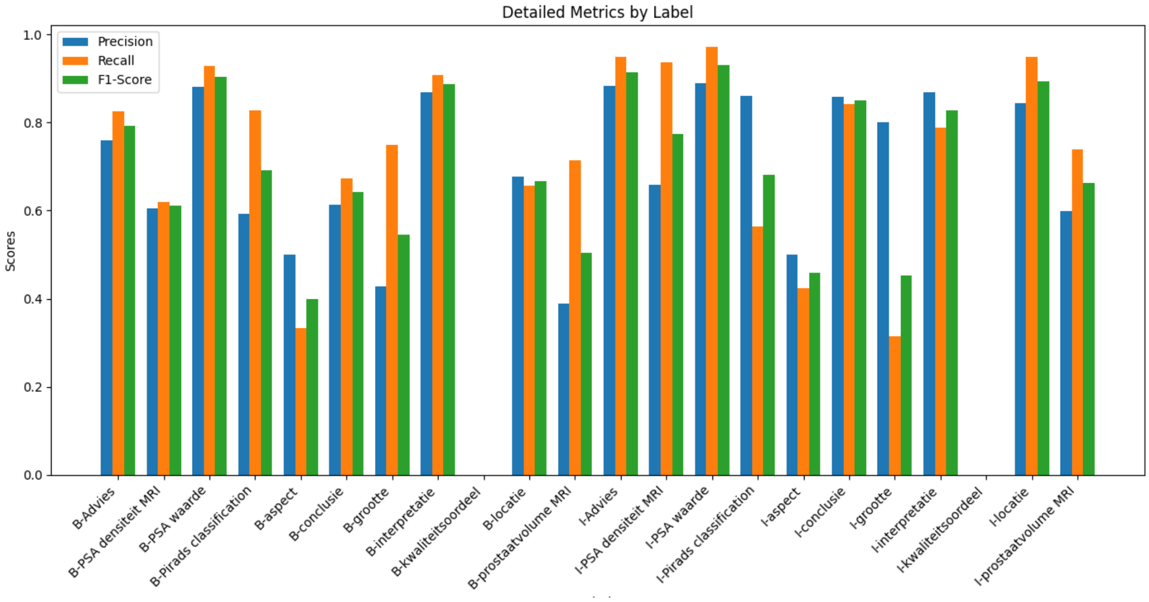


Figure 4.5: Hybrid-CRF Model - Detailed Evaluation Metrics

The highest performance was achieved for field PSA Waarde with an F1-score of 0.94. The lowest F1-score was for the field aspect with 0.45 excluding the rare ”kwaliteisoordeel” field which was present in only two reports. The next least occurring field was ”aspect”, which can be the reason for low performance on its identification. This suggests that under-represented fields in the training data lead to lower performance, highlighting the need for more representative data.

2. Qualitative Analysis

   This approach evaluated the model’s practical utility in a real-world setting. We developed a small interface that displays missing fields and compliance issues(see figure 4.9) for this purpose. Radiology reports were transferred from the reporting interface to our server via JSON files. The radiologist who used this model found it easy as it just shows the missing fields almost correctly but found that some fields were not identified even after reporting them, which made him look again through his report in vain. To understand which types of fields were not identified correctly the qualitative analysis of this reporting was done. It was found that if the reporting structure is not consistent with the structure of the report that was given for training, the model fails to identify the fields, resulting in wrong predictions.

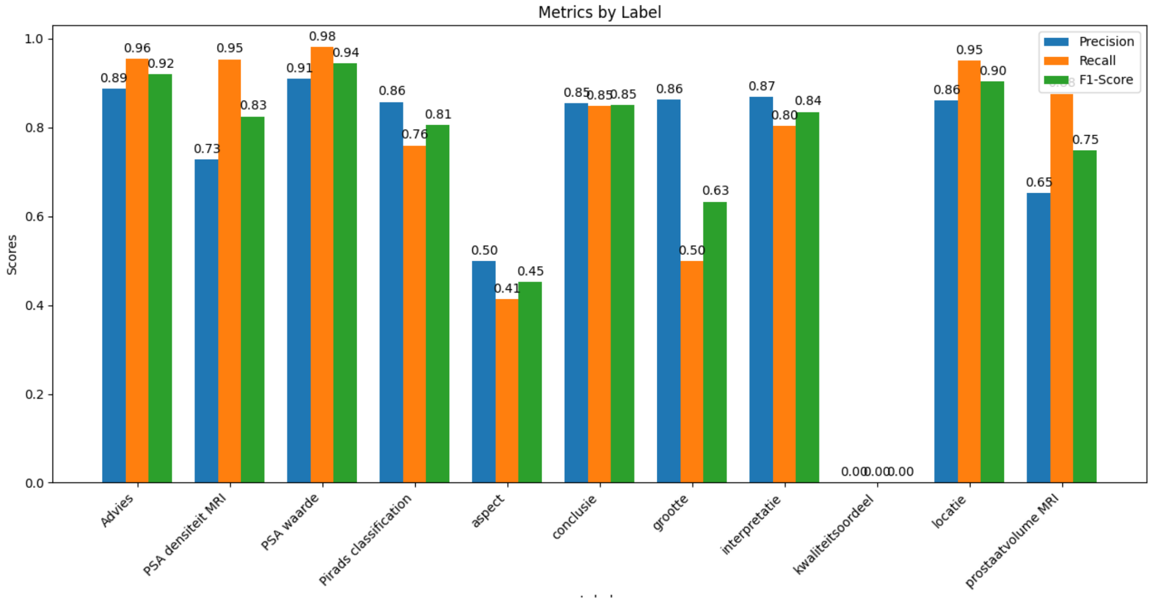


Figure 4.6: Hybrid-CRF Model - Combined Evaluation Metrics

Among the approaches evaluated, the Hybrid CRF model demonstrated the most promising performance in identifying the fields. Implementing a non-interactive interface allowed us to test the model's efficacy with real-time reporting. This setup was able to identify missing fields and ensure compliance with predefined rules.

However, the challenge with this approach is its limitation in handling new types of reporting structures. The training data given was more structured explicitly marking abnormalities and corresponding fields. The new set of evaluation data that is used for the interface evaluation lacked such detailed structuring Additionally, the model struggled with identifying underrepresented fields such as ”aspect” and ”kwaliteisoordeel.”

The language models on the other hand exhibited overfitting due to the limited annotated data. This lack of sufficient data led to difficulties in feature identification and generalization.

To address these limitations, we need to explore strategies to enhance data representation and model generalization. One potential solution is data augmentation, which can help create a more diverse training dataset. However, in the medical domain, where fabricated data may not be applicable, having real data is important. Therefore, we propose developing a solution to improve the efficiency of annotated data collection. The next chapter will discuss the methodology for this and the integration of the models into a more interactive interface.

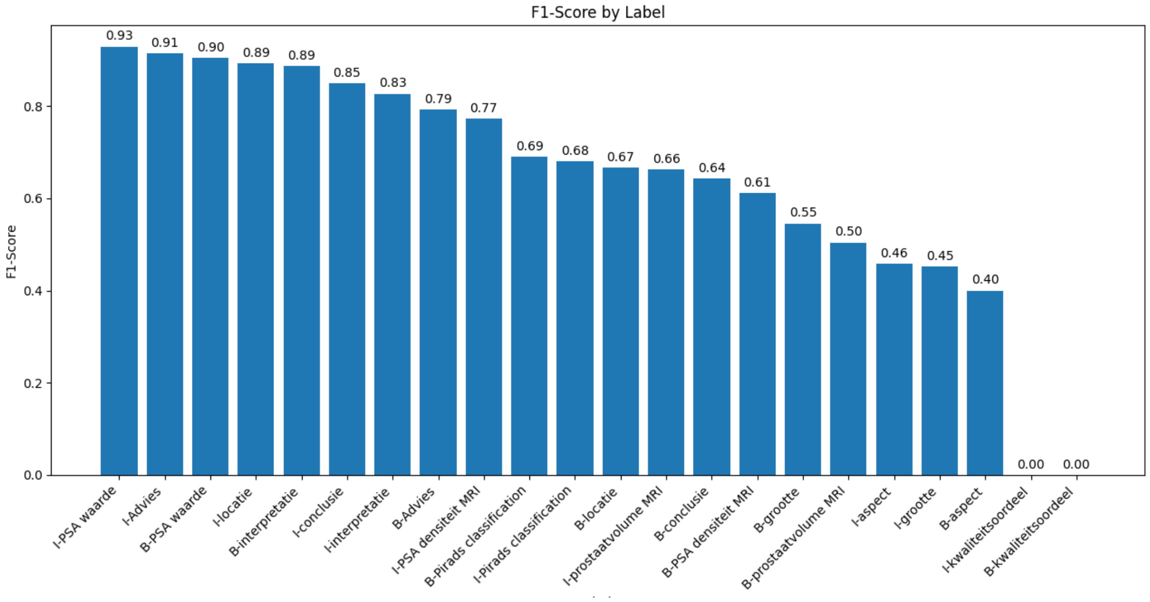


Figure 4.7: Hybrid CRF- F1 score Detailed Performance

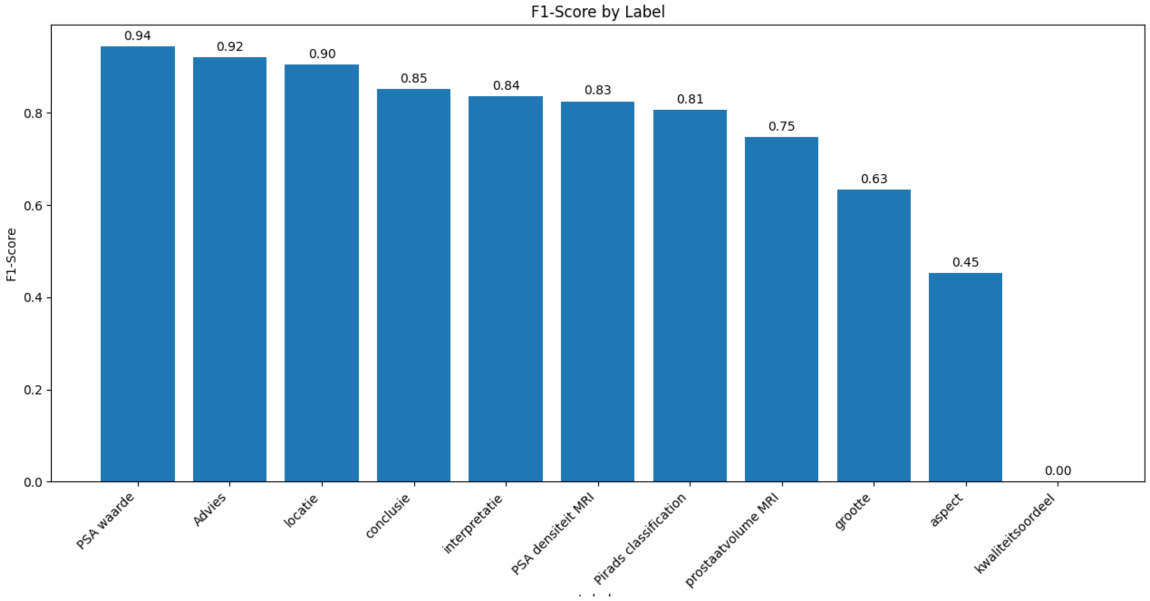
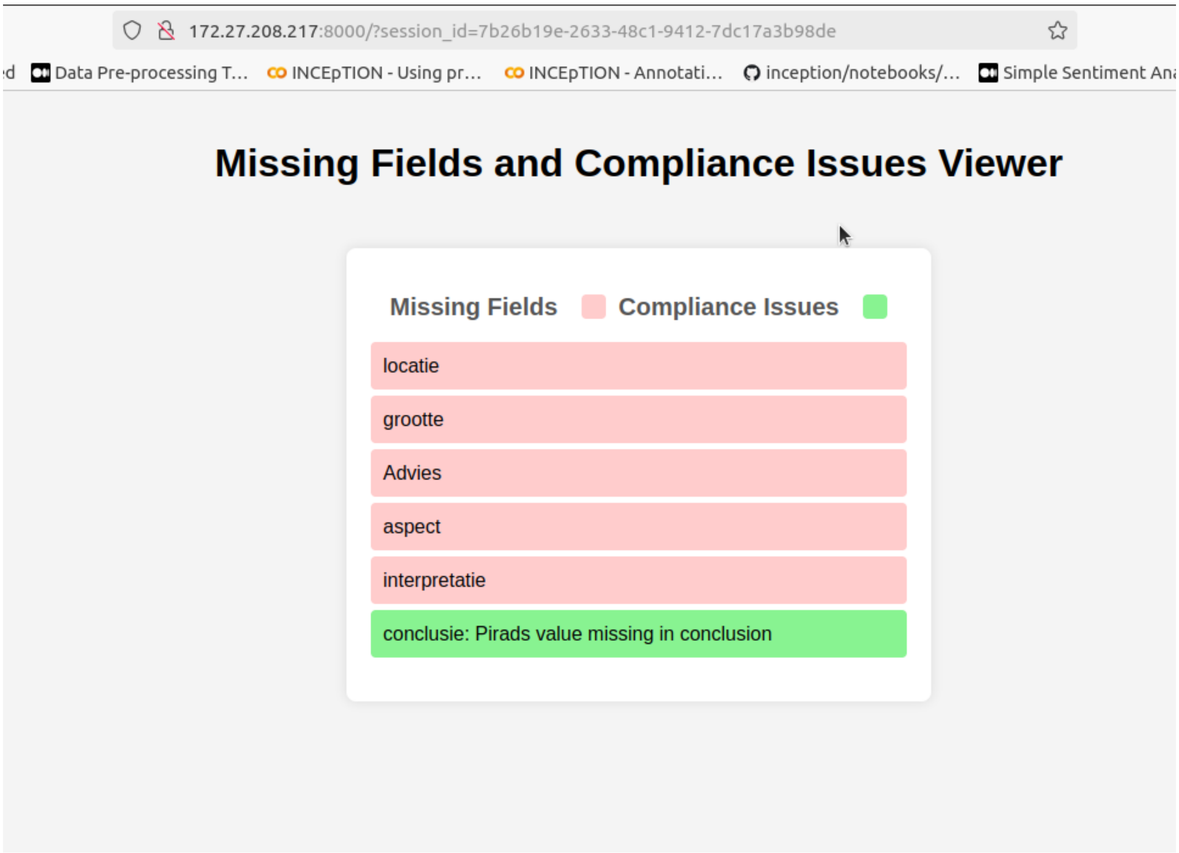


Figure 4.8: Hybrid CRF- F1 score Combined Performance



Missing Fields and Compliance Issues Viewer

Missing Fields 🟥 Compliance Issues 🟩

- locatie
- grootte
- Advies
- aspect
- interpretatie
- conclusie: Pirads value missing in conclusion

Figure 4.9: Interface showing missing fields and Compliance issue

Chapter 5

Interface and Annotation

The previous chapter described how the model was developed to ensure the completeness and compliance of the reports. However, to leverage the model in real-time, it is crucial to integrate it seamlessly into the radiologists' workflow.

Another important fact is the limited availability of annotated datasets we have. Annotating data in the medical domain is particularly challenging as it typically requires the involvement of hospital staff. In our case, the radiologists were the annotators, which required them to allocate time from their already busy schedules. As highlighted in the previous chapter, the limited annotated data has been a critical issue. Additionally, reporting patterns evolve, making it essential to continuously update the dataset with new annotations.

To mitigate this challenge, we propose developing a solution that integrates data annotation within the reporting workflow. This approach would enable the continuous collection of annotated data throughout the reporting process, leading to a richer and more representative training dataset, ultimately enhancing model performance.

Various methods can help reduce the reliance on annotated data, but the most effective solution would be one that facilitates easier data collection. Thus we decided to build an interface that not only integrates our model but also functions as an annotation platform. This enables radiologists to annotate data while performing their regular reporting tasks. This chapter discusses the journey of developing an interface, the evaluation of it and the challenges we encountered during the process, addressing our second research question.

The chapter is organised into three sections:

Section 5.1: Methodology– Developing the Interface

Section 5.2: Evaluation of Interface and Result Analysis

Section 5.3: Extraction of Annotations from User Interactions

5.1 Methodology– Developing the Interactive Interface

The Interface we are developing will be able to show the output of the model, and also be able to annotate the report which can be further used for retraining the model or for other future tasks. Before developing the interface, the functionality and requirements for the interface were discussed.

5.1.1 Requirements for the Interactive Interface

Before developing the interface, a thorough analysis of the potential requirements and functionalities of the interface was conducted. This helped in designing and developing the interface and its functionality such that it integrates seamlessly with the radiologist's workflow and enhances the efficiency of the annotation process and model's application.

1. **Integration with Radiologists' Workflow**
   Integrating the model into the radiologists reporting flow is important as without this the model and the interface developed cannot be used in the practical world. This involves designing an interface that is easily accessible and works without disrupting their usual workflow. This will include how the interface should be integrated into their reporting workflow.

   To design the interface the endpoint where the user needs the functionalities of the interface has to be known. After the initial discussion with the radiologist, it was determined that the interface should be incorporated at the end of the report preparation process. Here, a "check" button could be introduced, enabling the radiologists to access the interface's functionalities after completing their reports.

2. **Visualization of Model Results**
   Another important functionality of the interface is to visualize the results of our model in a clear and interpretable way. So the results the radiologists need to see are the missing fields and compliance issues in the report so that they can update it. This also includes how they can interact with the results, like how they can add the missing fields and how they can correct the compliance issues.

3. **Efficient Data Collection from User Interactions**
   In addition to showing the model results, using the user interactions for collecting data is a significant functionality. This means that this interface can also be used as an annotation tool that provides data for model retraining and other future tasks.

To do this interface should enable the radiologists to provide their feedback directly through their interactions. These functionalities include allowing users to correct the model predictions. To get more information, not only the missing fields but also the identified fields by the model can be highlighted and labelled, so that the user can correct it if wrong. More details on these functionalities are discussed in the coming section 5.1.2.

These functionalities should be designed in such a way that they shouldn't take much time and disrupt the radiologists' workflow. Interactions need to be intuitive, making it easy for radiologists to use the interface efficiently while maximizing the extraction of information from their actions.

4. **Extraction of Annotations from User Logs**

   The user interactions can be stored as logs and the annotations can be extracted from these user logs. These logs capture detailed information about how radiologists interact with the interface including any corrections or additions to the model's output. To extract the annotations in a format suitable for retraining the model, specialized code needs to be developed. This code will parse the logs, identify relevant annotations, and structure the data in a way that is compatible with the model's retraining pipeline.

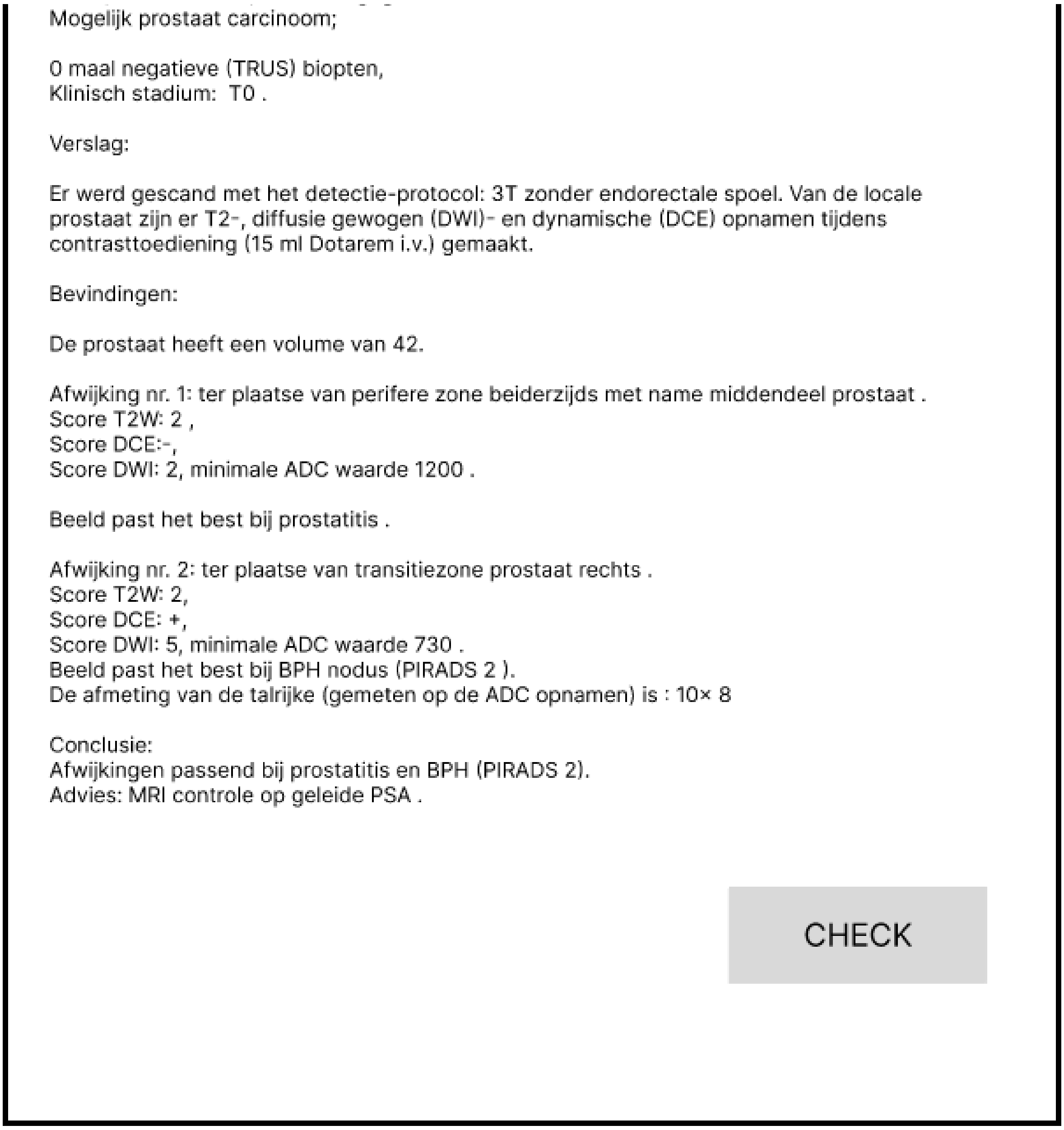
5.1.2 Prototype of the Interactive Interface

After identifying the use cases, a prototype of the interface was developed. The prototype shows how the interactions work and how the model results are visualised. The prototype was developed with Figma [38]. Figures 5.1, 5.2, 5.3, 5.4, and 5.5 show the interactions in the prototype.

After the report is ready, there will be a check button as shown in figure 5.1 which will take the user to the page where the report with the following interactions is seen.

1. **General Interactions**

   After clicking the check button, the user will be navigated to the page where the report with the identified fields and missing fields is shown. The general interactions on this page are the undo and redo functionality which the user can use to go back and forward to their interactions. There is also a suggestion box side to the report that shows the missing fields and the compliance guidelines. A ”Save” button is also available to save the reports after the checking. Figure 5.2 shows the prototype after clicking the check button. The specific interactions on this page are defined next.



Mogelijk prostaat carcinoom;

0 maal negatieve (TRUS) biopten,
Klinisch stadium: T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel. Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een volume van 42.

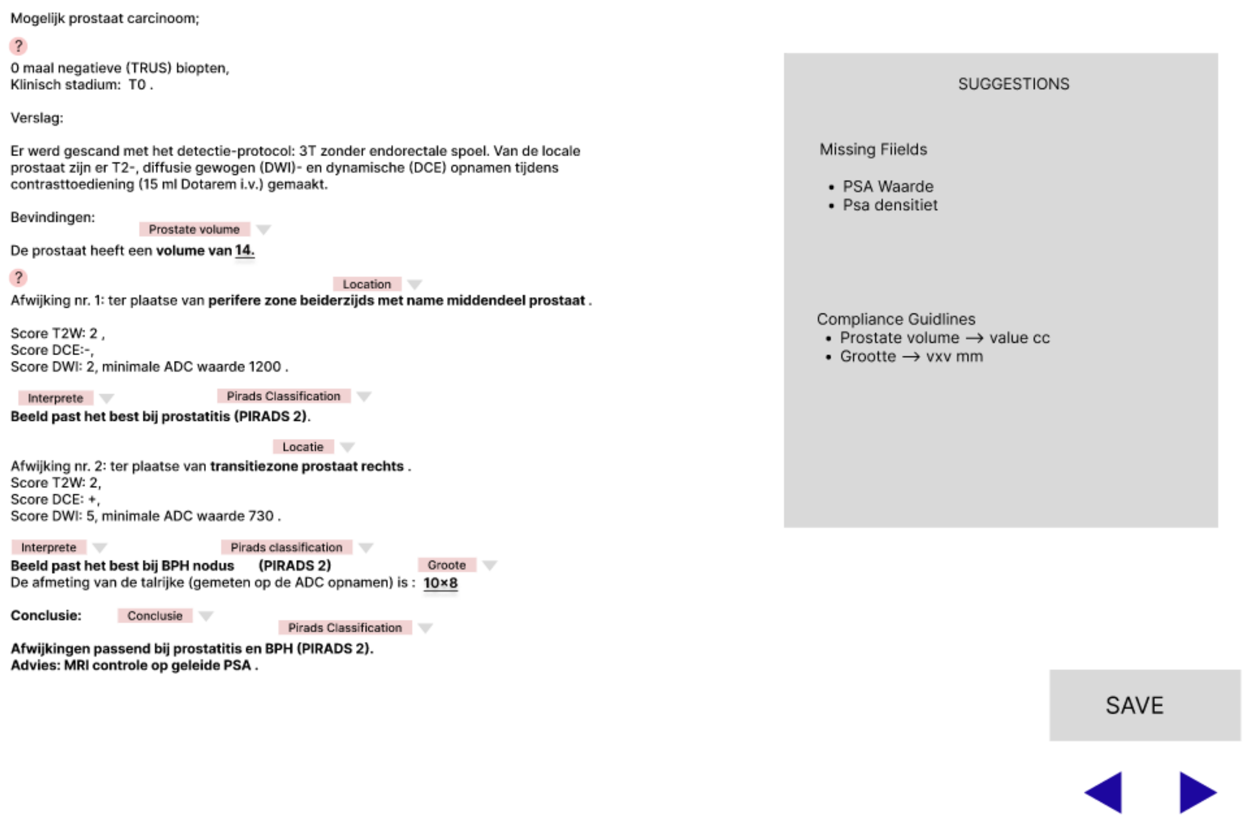
Afwijking nr. 1: ter plaatse van perifere zone beiderzijds met name middendeel prostaat .
Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .

Beeld past het best bij prostatitis .

Afwijking nr. 2: ter plaatse van transitiezone prostaat rechts .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .
Beeld past het best bij BPH nodus (PIRADS 2).
De afmeting van de talrijke (gemeten op de ADC opnamen) is : 10× 8

Conclusie:
Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .

Figure 5.1: Prototype- Check Page



Mogelijk prostaat carcinoom;

?

0 maal negatieve (TRUS) biopten,
Klinisch stadium: T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel. Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een **volume van 14.**

?

Afwijking nr. 1: ter plaatse van **perifere zone beiderzijds met name middendeel prostaat** .

Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .

Beeld past het best bij prostatitis (PIRADS 2).

Afwijking nr. 2: ter plaatse van **transitiezone prostaat rechts** .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .

Beeld past het best bij BPH nodus (PIRADS 2)
De afmeting van de talrijke (gemeten op de ADC opnamen) is : **10×8**

Conclusie:

Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .

SUGGESTIONS

Missing Fiields

- PSA Waarde
- Psa densitiet

Compliance Guidlines

- Prostate volume → value cc
- Grootte → vxv mm

Figure 5.2: Prototype- 2nd Page

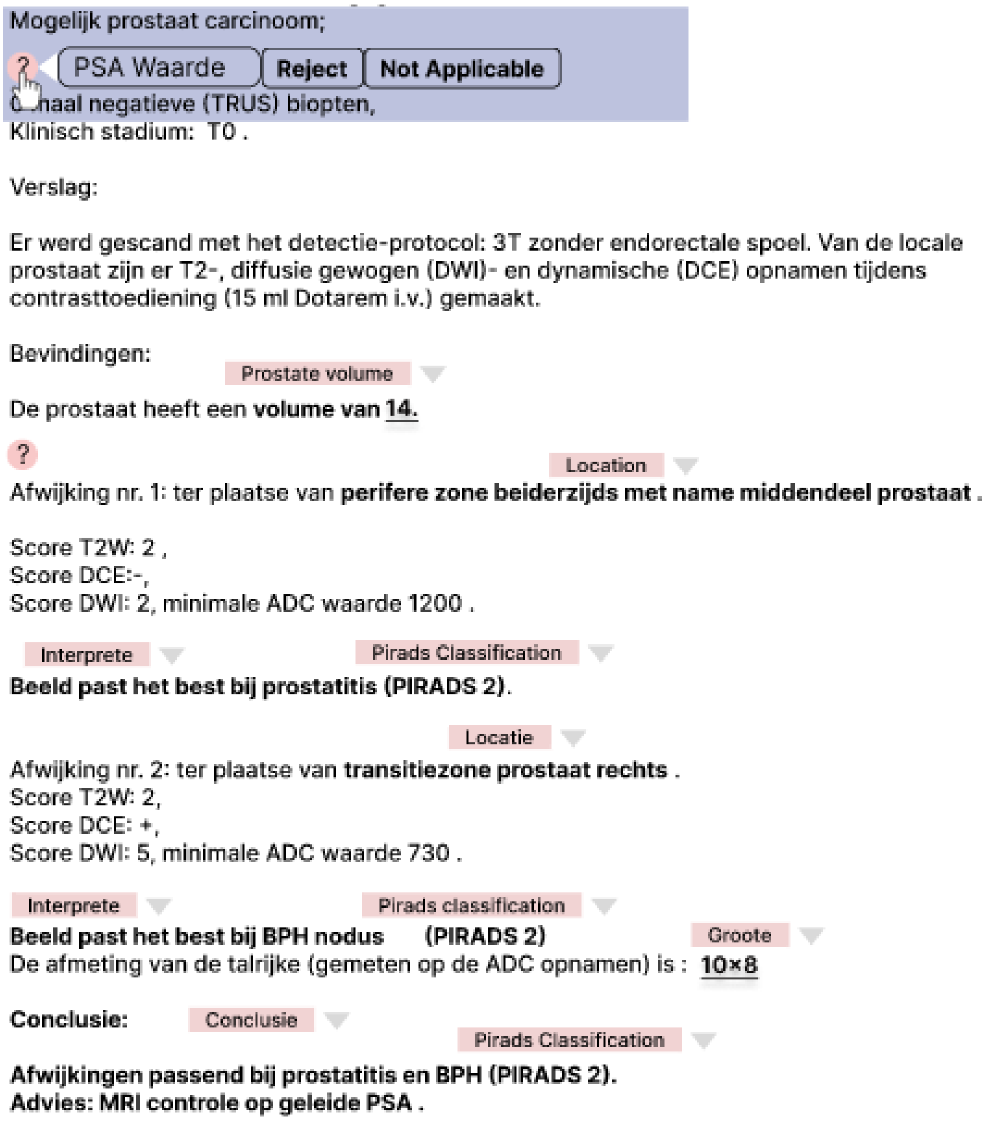
2. **Missing Fields**

   Missing fields are displayed as the '?' symbol in the report. Clicking on them shows the missing field and two other options for rejecting this suggestion. On clicking on the missing field label, the user can enter the field value. The two options for rejection are to distinguish the reason for rejection.

   The user may reject a suggestion of a missing field, because it is already present, in this case, the button "REJECT" is given. The second reason may be the missing field does not apply to the particular report, like "Advies" which is not always provided in a report. For this case, the button "NOT APPLICABLE" is given. The reason for distinguishing between the two is that the model can learn the difference and predict the missing fields accordingly. Figure 5.3 shows the interactions associated with the missing fields, in this example, the PSA Waarde is the missing field.

3. **Compliance Issues**

   The Compliance issues are shown as underlined text, when hovered over it it shows the right format. By clicking on the right format the text will be replaced with it. There is also an option "Reject" for rejecting this suggestion. Figure 5.4 shows this interaction, in this case, the prostate volume 14 is not in the right



Mogelijk prostaat carcinoom;

naal negatieve (TRUS) biopten,
Klinisch stadium: T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel. Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een **volume van 14.**

Afwijking nr. 1: ter plaatse van **perifere zone beiderzijds met name middendeel prostaat** .

Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .

Beeld past het best bij prostatitis (PIRADS 2).

Afwijking nr. 2: ter plaatse van **transitiezone prostaat rechts** .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .

Beeld past het best bij BPH nodus (PIRADS 2)
De afmeting van de talrijke (gemeten op de ADC opnamen) is : 10×8

Conclusie:

Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .

Figure 5.3: Prototype- showing Missing Fields Interaction

format it should be 14 cc.

4. **Identified Fields**

   The Hybrid CRF field model also identifies the fields in the report. This is shown by highlighting those texts and the corresponding label is marked at the top. The label is associated with a drop-down menu of other labels. The user can change the label if the model predicts it wrongly by clicking on the labels in the drop-down list. Figure 5.5 shows this interaction, in this example, interpretatie is the label and the drop-down menu can be seen while clicking on it.

The prototypes were developed to visualize the interactions. As mentioned above each interaction satisfies the use cases of the interface, and the next step is to implement this interface and evaluate it.

5.1.3 Implementation of the Interface

After developing the prototypes and designing the interactions, the next step was to implement the interface to put these concepts into practice. This includes integrating the NLP model with a user-friendly front end that allows the users to interact with the system efficiently.

Software and Frameworks

The interface was developed using the following packages and technologies to meet the back-end and front-end requirements:

Flask(Python)

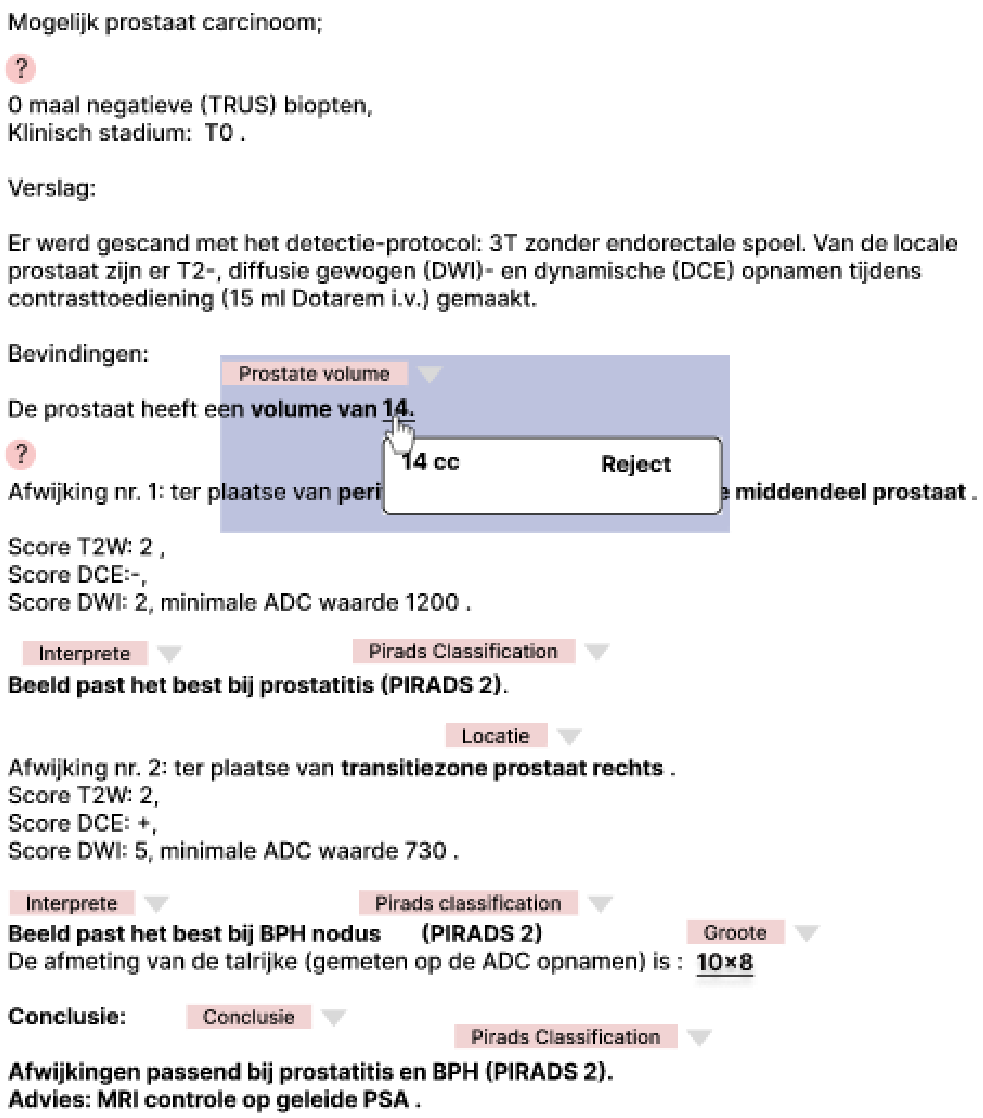
Flask [39] is chosen for the back-end to integrate the model into the interface because of its simplicity and effectiveness. Flask is a small Python framework that has a set of helpful features and tools that ease the creation of web applications in Python. It handles requests, processes the NLP model results, and communicates the results back to the front end.

HTML, CSS, Javascript

These web technologies were used to create the front-end interface. HTML creates the structure of the interface, CSS adds the style to the structure, and Javascript adds the interactive elements and functionality to the interface. This combination allows to build an interactive interface with the required functionalities.

Gunicorn

Gunicorn is a Python WSGI (Web Server Gateway Interface) server which serves the Flask application. It is a robust and efficient server and can handle multiple



Mogelijk prostaat carcinoom;

0 maal negatieve (TRUS) biopten,
Klinisch stadium: T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel. Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een **volume van 14.**

Afwijking nr. 1: ter plaatse van **peri** **middendeel prostaat** .

Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .

Beeld past het best bij prostatitis (PIRADS 2).

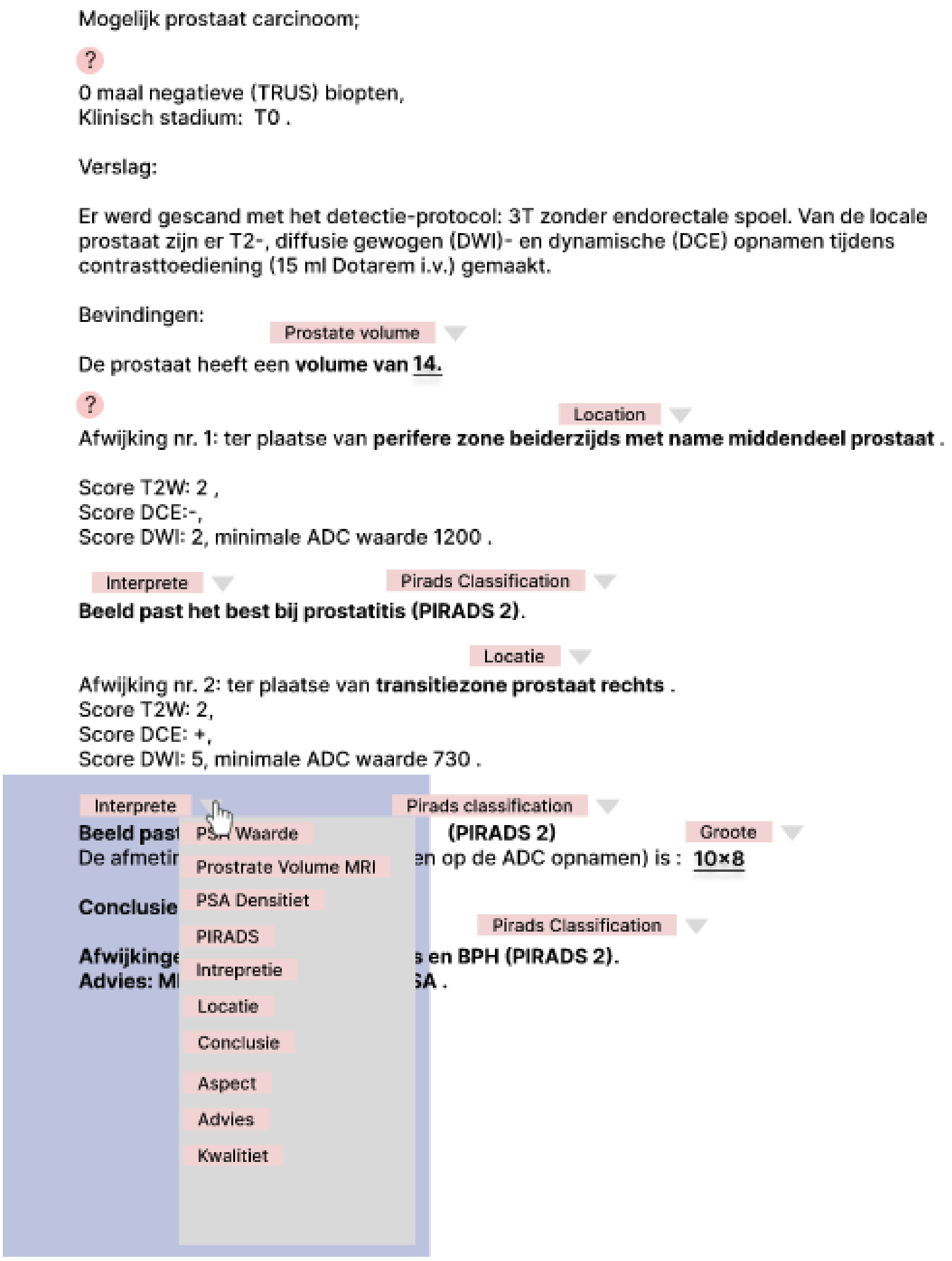
Afwijking nr. 2: ter plaatse van **transitiezone prostaat rechts** .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .

Beeld past het best bij BPH nodus (PIRADS 2)
De afmeting van de talrijke (gemeten op de ADC opnamen) is : **10×8**

Conclusie:

Afwijkingen passend bij prostatitis en BPH (PIRADS 2).
Advies: MRI controle op geleide PSA .

Figure 5.4: Prototype- showing Compliance Issue Interaction



Mogelijk prostaat carcinoom;

?

0 maal negatieve (TRUS) biopten,
Klinisch stadium: T0 .

Verslag:

Er werd gescand met het detectie-protocol: 3T zonder endorectale spoel. Van de locale prostaat zijn er T2-, diffusie gewogen (DWI)- en dynamische (DCE) opnamen tijdens contrasttoediening (15 ml Dotarem i.v.) gemaakt.

Bevindingen:

De prostaat heeft een **volume van 14.**

?

Afwijking nr. 1: ter plaatse van **perifere zone beiderzijds met name middendeel prostaat** .

Score T2W: 2 ,
Score DCE:-,
Score DWI: 2, minimale ADC waarde 1200 .

Beeld past het best bij prostatitis (PIRADS 2).

Afwijking nr. 2: ter plaatse van **transitiezone prostaat rechts** .
Score T2W: 2,
Score DCE: +,
Score DWI: 5, minimale ADC waarde 730 .

Beeld past **(PIRADS 2)**
De afmetir en op de ADC opnamen) is : 10×8

Conclusie

Afwijkinge s en BPH (PIRADS 2).
Advies: MI SA .

Figure 5.5: Prototype- showing Identified Fields and the Interactions associated with them

requests. Gunicorn is easy to implement and thus was chosen as the server to host our interface.

Interface Architecture

The architecture of the interface was developed to ensure smooth communication between the front end and back end. The current system used for reporting at the hospital is a speech-to-text system. So initial idea was to integrate the interface into that system, but for the time being it isn't possible, so the interface was implemented such that the radiologist can upload the reports and check. This is not a convenient approach, but for the evaluation, this was a solution we took.

When the report is uploaded it is sent to the flask application where the trained hybrid CRF model is loaded. The model processes the report and provides the result, which is then returned to the front-end interface with the interactions and functionalities discussed in the above section 5.1.2.

The interface has an upload page, with instructions to ease the users to interact 5.6. The check button at the upload page takes them to the page where the results of identified fields and missing fields in the reports will be shown as in figure 5.7. On the same page where the results are shown, there is a model performance box, where the users can see the performance of the model through the evaluation metrics- Precision, recall, F1 score (see Figure 5.7). This can be helpful for the radiologists to understand how capable is the model in identifying the fields.

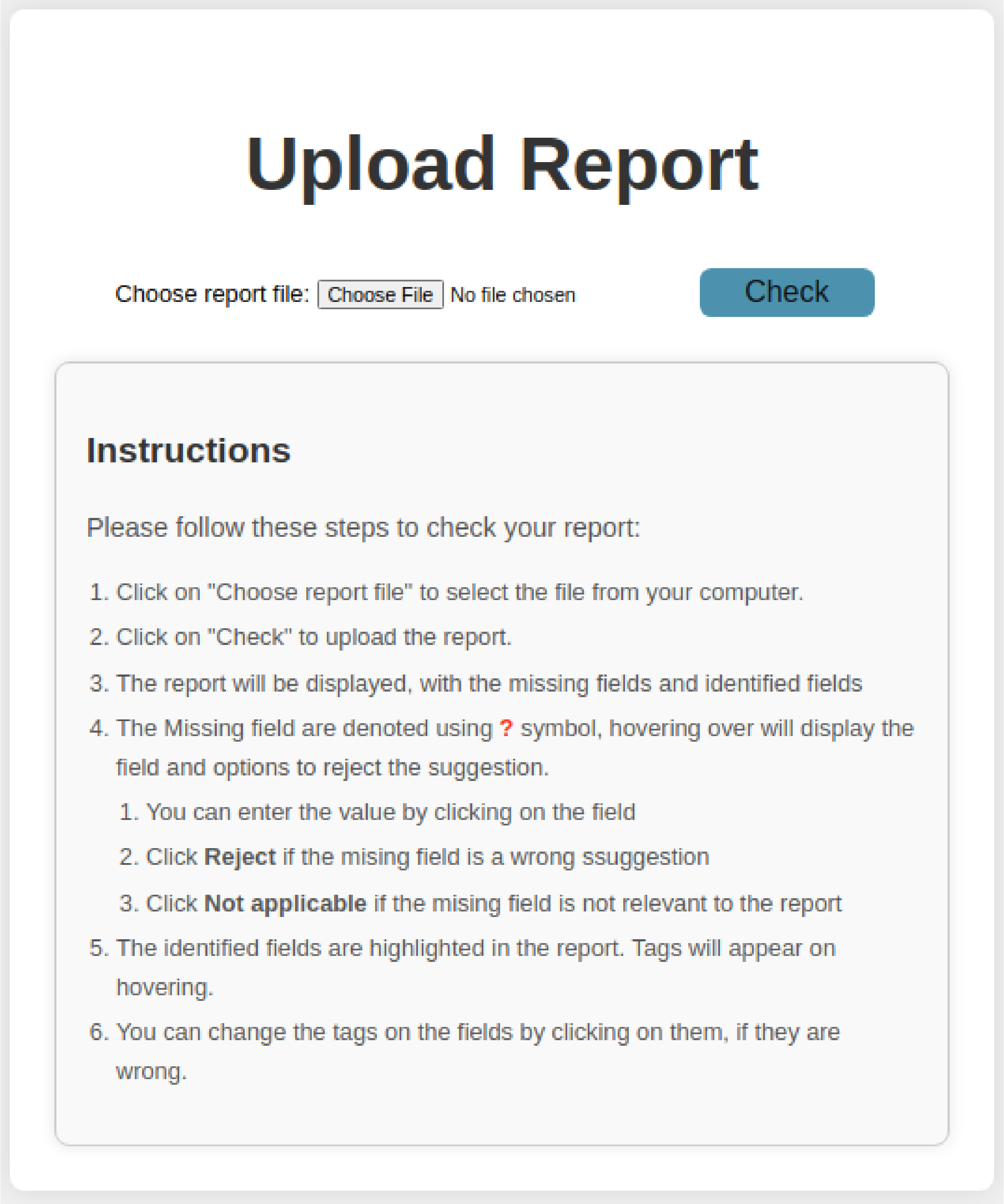
The other Interactions are the same as described in the prototype and the implementation of those in the interface is shown in the figure 5.8. Additional thoughts were put in the colouring scheme of the rejection buttons, red for rejecting and green for not applicable as red symbolises wrong and green symbolises right (but here the field is not applicable).

Log Messages for User Interactions

The interactions of the user have to be collected to extract annotations from them. For this purpose carefully analyse log messages where included in the interface. The extraction of annotation from the interactions is done by developing codes. The following information was used to create the annotation.

1. Missing Fields Interactions

   When the user selects the missing field to enter the field value, the logs get created and from that the field label and the value can be extracted to add to the identified fields. If the user rejects the suggestion by selecting ”REJECT”, that will be removed from the missing field predictions, and a recheck of the

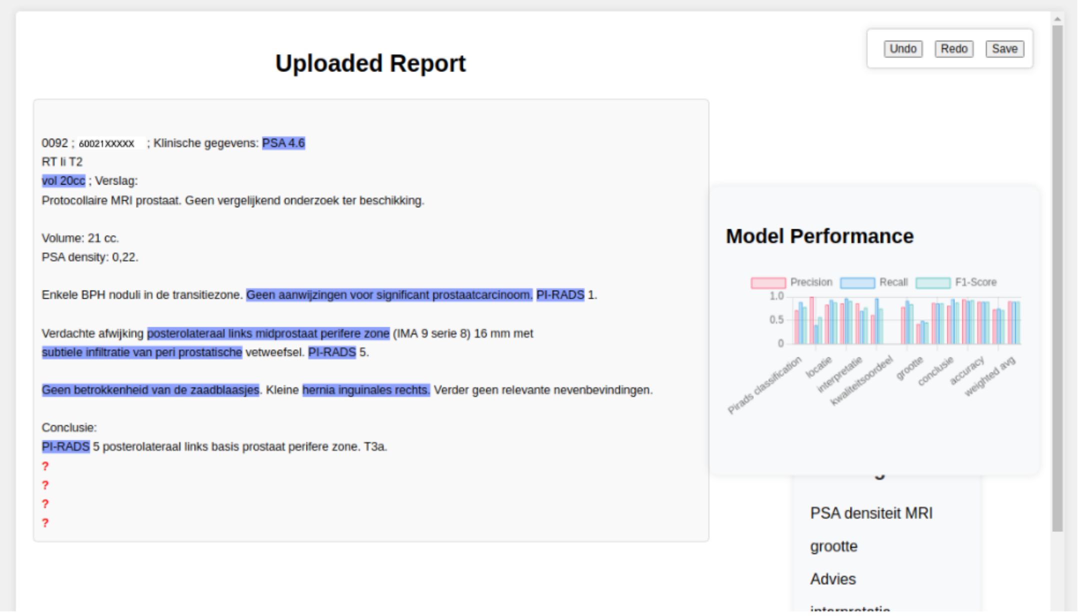


Instructions

Please follow these steps to check your report:

1. Click on "Choose report file" to select the file from your computer.
2. Click on "Check" to upload the report.
3. The report will be displayed, with the missing fields and identified fields
4. The Missing field are denoted using ? symbol, hovering over will display the field and options to reject the suggestion.
    1. You can enter the value by clicking on the field
    2. Click **Reject** if the mising field is a wrong ssuggestion
    3. Click **Not applicable** if the mising field is not relevant to the report
5. The identified fields are highlighted in the report. Tags will appear on hovering.
6. You can change the tags on the fields by clicking on them, if they are wrong.

Figure 5.6: Upload Page of Interface



Uploaded Report

0092 ; 60021XXXXX ; Klinische gegevens: PSA 4.6
RT li T2
vol 20cc ; Verslag:
Protocollaire MRI prostaat. Geen vergelijkend onderzoek ter beschikking.

Volume: 21 cc.
PSA density: 0,22.

Enkele BPH noduli in de transitiezone. Geen aanwijzingen voor significant prostaatcarcinoom. PI-RADS 1.

Verdachte afwijking posterolateraal links midprostaat perifere zone (IMA 9 serie 8) 16 mm met
subtiele infiltratie van peri prostatische vetweefsel. PI-RADS 5.

Geen betrokkenheid van de zaadblaasjes. Kleine hernia inguinales rechts. Verder geen relevante nevenbevindingen.

Conclusie:
PI-RADS 5 posterolateraal links basis prostaat perifere zone. T3a.
?
?
?
?

Model Performance

Figure 5.7: Interface Page showing the Results

identified fields happen. If ”NOT APPLICABLE” is selected that will be still marked as a missing field with additional info for marking it wasn’t right for the particular report. By doing this maybe with more data the model will be able to identify a pattern of why a particular field is not applicable based on the other information in the report.

2. Identified Fields Interactions

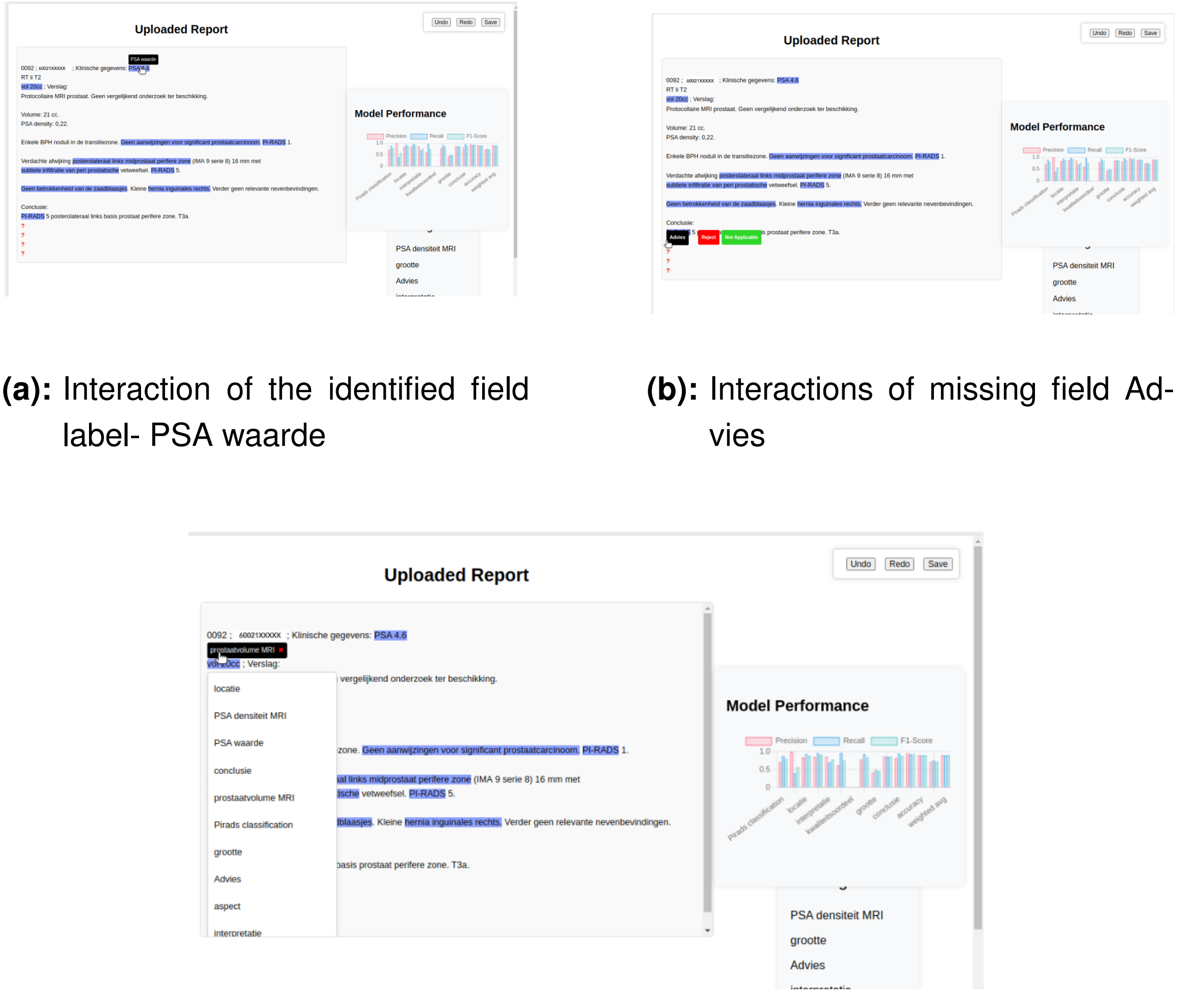
   The user can select a different label for the identified fields if the predicted label by the model goes wrong. This interaction can also be logged and can be used to change the labels for the annotation. Also if a label is deleted by the user, then the identified field can be deleted from the set.

Once the interface was developed, the next step was to evaluate its effectiveness.

5.2 Evaluation and Result Analysis

5.2.1 Evaluation set up

The evaluation of the interface was done with the help of two radiologists from ZGT. For this process, new reports were arranged from the hospital. The evaluation setup



Uploaded Report

0092 ; 60021XXXXX ; Klinische gegevens: PSA 4.6
RT li T2
vol 20cc ; Verslag:
Protocollaire MRI prostaat. Geen vergelijkend onderzoek ter beschikking.

Volume: 21 cc.
PSA density: 0,22.

Enkele BPH noduli in de transitiezone. Geen aanwijzingen voor significant prostaatcarcinoom. PI-RADS 1.

Verdachte afwijking posterolateraal links midprostaat perifere zone (IMA 9 serie 8) 16 mm met subtiele infiltratie van peri prostatische vetweefsel. PI-RADS 5.

Geen betrokkenheid van de zaadblaasjes. Kleine hernia inguinales rechts. Verder geen relevante nevenbevindingen.

Conclusie:
PI-RADS 5 posterolateraal links basis prostaat perifere zone. T3a.

(a): Interaction of the identified field label- PSA waarde

(b): Interactions of missing field Advies

(c): Dropdown menu showing the labels

Figure 5.8: Interactions available in the Interface

involved asking the radiologists to interact with the interface and assess its functionality. Given the time constraint of the radiologists, one of them evaluated 20 reports while the other evaluated 10 reports. After their interaction with the interface, they were asked a series of questions to determine if the interface met our expectations. The questions were as follows:

1. How was the experience with the interface?
2. How would you rate the ease of use of the interface?
3. How would you describe the time required for interaction with the interface?
4. Does incorporating this into the workflow make the reporting easier?

5. How does the annotation process compare to other methods you've used?

6. What suggestions do you have for improving the interface?

7. Where would you prefer the interface functionalities to be located or integrated?

5.2.2 User feedback and Findings

1. **User Experience with the Interface**

   On observing the interactions, initially, radiologists took some time to understand the functionalities. However, as they continued using it, the interaction became more familiar and intuitive for them. When asked about the same, they said that with increased usage, they grew more comfortable with the functionalities. This feedback indicates that there is a positive learning curve where initial challenges can be mitigated with familiarity, leading to a smooth experience.

2. **Ease of Use of the Interface**

   Aside from the initial confusion, particularly with rejection options of the missing fields, the radiologists found the interface easy to use once they understood its purpose. The functionalities of just clicking for changing and adding labels made the process easier. The observation from the feedback is that the interface is user-friendly and intuitive.

3. **Time Required for Interaction**

   As mentioned before, initially learning about the interaction was time-consuming, but the subsequent interactions were quicker. When asked about the same, one of them mentioned that the colour-coding of the buttons (red and green) for rejection helped make decisions more rapidly based on visual cues. However, the other radiologist mentioned that he was confused with the button labels especially the term ”REJECT”, which took some time to understand. Thus changing the term ”REJECT” to a more intuitive term like ”WRONG” would be easier to comprehend, as this better reflects the context of correcting an incorrect model prediction.

4. **Impact on Reporting Workflow**

   One of our main goals was to make the integration of the model and annotation work along with the reporting workflow. The radiologists mentioned that integrating this interface and functionalities directly into their workflow would be beneficial. They also expressed the concern that frequent model errors could disrupt their workflow, showing the requirement of high accuracy in the model suggestions.

5. **Comparison of Annotation Process**

   Another key use case of this interface is the extraction of annotations from user interactions. Generally, including for this research radiologist perform the annotations using a separate platform, which takes additional time outside of their regular duties. In contrast, annotation through the interface is performed while checking the report, seamlessly integrating the process into their existing workflow. Thus this annotation process is seen as a more efficient method as it eliminates the extra task of annotation. The radiologists also appreciated the time-saving aspect of the interface for annotation and it is a better approach than the other. But they also said that it would be good to have the option to select and label the text directly in the report, as this feature would provide them more freedom to interact and correct the model.

6. **Integration of interface Functionalities**

   Currently, for the evaluation purpose the interface was used at the end of the reporting, that is after completing the whole report. To understand if this is effective or if the integration of the interface functionalities from the beginning of the reporting process is effective, we sought the radiologists' opinions. They indicated that having the interface available through the reporting process would better support their workflow. However, the current system they are using limitations pose a challenge to this integration.

7. **Suggestions for Improvement**

   Finally, after the interactions, we asked the radiologists for suggestions to improve the interface. One of them was to change the term ”REJECT” to a term that relates more to the action. Since the action here is to avoid the suggestions of missing fields as it is wrongly predicted by the model, we concluded that the term ”WRONG” would be more suitable.

   Another suggestion was to incorporate a selection and label interaction where the users can select and label a text in the report if the model fails to. This is a useful interaction and can be used for annotation extraction.

8. **Evaluation Data and patterns**

   For the evaluation of the interface, a dataset of 100 reports was used. However, these reports differed in structure compared to the training data. The reports in the evaluation set were less structured when compared to the training data, with no separate sections for each abnormality. The field prostrate volume appeared twice with different values. Upon inquiry, the radiologists explained that reports might include volume measurements from the ultrasound, but the volume measured during reporting is considered the final value. Also, some reports used millilitres (ml) instead of cubic centimetres (cc), which is the standard metric according to guidelines.

While observing the radiologists' interactions with the interface, it was noticed that some of the missing field suggestions were rejected with a pattern. To clarify we asked the radiologist, who explained that when the PIRADS value is less than 3 (PIRADS 1 and PIRADS 2), fields such as size, location, aspect, and interpretation are not necessary and can be excluded as missing fields. We also found that the reporting pattern changes, with time and radiologists, so more the reports we collect more the data can be provided to the model, and the model can adapt to the task. Thus collecting annotation while using the model seems to be a good approach.

So after the evaluation, a small amount of data was collected as a result of the user interactions. The next section outlines the method of obtaining annotation from these user interactions.

5.3 Extraction of Annotation from User Interactions

The development of the interface has two goals- to integrate the model and to collect annotations from their interactions. This section outlines the process of extracting annotations from user interactions and converting them into a format suitable to use as training data for the model.

To capture the user interactions, log messages were added to the interface implementation. For each action, a specific log message was formatted.

After preparing the log messages, the next step was to extract the annotations from these messages. The evaluation results were analyzed, and a code was developed to systematically extract the necessary information from the log messages, transforming them into a format that could be used for model retraining. Table 5.1 illustrates the interactions, their corresponding log messages, and the annotation formats extracted from them.

Due to time limitations in collecting enough data for retraining, this research does not include model retraining with the newly collected annotations. However, the following steps are recommended to continue using this approach to enhance model performance:

Interaction Type	**Log Message**	**Corresponding Annotation**
Identified Fields	`Identified entities: [(value1, covered_text1), (value2, covered_text2)]`	`{'Value': value1, 'CoveredText': covered_text1}, etc.`
Missing Fields- Addition	`Entered value for <label> value: <value>`	`'Value': <label>, 'CoveredText': <value>`
Label removal	`Label "<label>" removed from "<text>"`	`Removes the entity with 'Value': <label>, 'CoveredText': <text>' from final annotations`
Text Selection	`Selected text "<text>" with label "<label>"`	`Adds 'Value': <label>, 'CoveredText': <text>' to final annotations`
Dropdown Change	`Dropdown Category Changed from <previous> to <new> for <item>`	`Updates the 'Value' of 'Covered Text' <item> from <previous> to <new>`

Table 5.1: Summary of Log Interactions and Corresponding Annotations

1. Extarct the annotations

   Using regular expression rules extract the annotation from the log messages as mentioned in the table 5.1.

2. Train the model

   Use the new set of annotations to train the model by incorporating them into the training set.

3. Evaluate the performance

   After training, evaluate the model's performance to assess improvements.

These steps can be integrated into a continuous pipeline allowing the model to improve automatically as radiologists interact with the system. The model's performance can be displayed in the interface, providing users with real-time visibility into how well the model is performing.

Chapter 6

Discussion and Recommendations

6.1 Discussion

In this research, we compared different models and chose a Hybrid Conditional Random Field model that can identify important fields in prostate screening radiology reports. The results were then used to predict the missing fields in the report thereby ensuring the completeness of the reports. Additionally, we created a set of regular expression rules to ensure that the reports adhered to compliance standards. We also explored other approaches such as Dutch Language Models but these models encountered overfitting issues due to the limited availability of annotated data. This led us to develop an innovative approach that integrates the model into the radiologists' workflow, allowing for both the use of the model in real-time reporting and the facilitation of the annotation process, making it more efficient.

The Hybrid CRF model demonstrated good performance in identifying the fields, that were well represented in the training data. However, it performed poorly in identifying underrepresented fields and failed to accurately identify fields when new reporting structures were introduced. The integration of an annotation process within the interface offers a solution to this limitation. By continuously collecting annotated data through the interface, we can retrain the model to adapt to new reporting structures and improve its performance in underrepresented fields.

Gaining more annotated data is also expected to help mitigate the overfitting issues observed with the language models, potentially enabling their application for the same task. The interactive interface has shown that annotation can be done along with the reporting. Although the interface is not yet fully integrated with the hospital's existing systems, its potential for future integration is promising. Radiologists also suggested that the interface could be even more useful if it provided real-time feedback during the reporting process, rather than only at the end. This would allow them to address any issues immediately, rather than reviewing the report at the end. However, for this to be effective, the model must achieve high accuracy in its predictions, as frequent incorrect predictions could disrupt the reporting process.

In summary, this research aimed to answer the following research questions and the findings are as follows:

”How accurately can NLP algorithms cross-check prostate screening reports with Guidelines to improve their completeness and compliance?”

RQ1: What models can be adapted to extract information from the report?

We tested different models including Dutch Language Models, Prompting, and Hybrid Conditional Random Fields to extract the information from the reports and identify them using the Named Entity Recognition (NER) approach. The Hybrid Conditional Random Field model was the effective model that identified the fields except for the fields which are underrepresented. This indicates that while the CRF model is adaptable, it requires more annotated data to improve accuracy in identifying all fields consistently.

RQ2: How can we ensure the compliance of the report?

The compliance of the report was ensured by defining regular expression rules, However, these rules applied only to fields with a defined structure, limiting their applicability to just three fields. Ensuring compliance for more complex, unstructured fields remains a challenge.

RQ3: What evaluation metrics should be used to measure the performance of the models?

To analyse the model performance we used quantitative and qualitative evaluation metrics. The quantitative evaluation of the Dutch Language models resulted in overfitting. The reason is considered due to the limited dataset. For quantitative evaluation, F1 scores were used to assess the accuracy of the CRF model, which ranged from 0.94 to 0.45, with the lower scores attributed to underrepresented fields like ”aspect” and ”grootte.” A qualitative evaluation was also performed for the CRF to see the performance in the real setting. This was done by implementing an interface that shows the results to the user while they are reporting. While the CRF model performed well in real settings, inconsistencies in reporting structures affected its performance.

Thus to address the first research question, the Hybrid CRF model identifies fields that are represented in the training dataset and can suggest the missing fields to the radiologists, which can help ensure completeness. The compliance as mentioned applies only to three fields and the regular expression rules can identify them. The identification of underrepresented fields and the new structure are the limitations of this approach. Increasing the annotated data for training may solve the problem of identification of underrepresented fields. The second research question of this study focused on the integration of the model and the improvement of the annotation process, with the following findings:

" How can models be integrated into the reporting workflow of radiologists to check the report and to improve the annotation process simultaneously?"

RQ1: What features should the user interface include to support seamless integration of NLP techniques into radiologists' workflow?
The interface was designed with features that supported seamless integration of the NLP model into the radiologists' workflow, without creating additional burdens. Prototypes were tested and refined to ensure that the interface was easy to use and helped radiologists without interrupting their workflow. Key features included highlighting missing fields and providing compliance checks after the report was completed.

RQ2: What features should the interface have, to adapt model learning from user interactions?
To adapt the model's learning, the interface allowed for interaction-based annotation collection. This was crucial, as it enabled the model to learn from radiologist interactions and continuously improve. Features that allow radiologists to correct the model's predictions were built into the interface, allowing real-time data annotation and collection.

RQ3: What metrics should be used to evaluate the effectiveness of NLP integration improving radiologists' reporting practices and annotation process?
The evaluation of the interface with the real users radiologists was conducted to understand whether the features of the interface integrate the model into the workflow and improve the annotation process. Although quantitative metrics were limited due to a small sample size, the qualitative results showed that radiologists appreciated the interface but requested real-time feedback during the reporting process. This suggests that future evaluations should incorporate more test cases to quantitatively measure improvements in reporting accuracy and time efficiency. The annotation extraction from the interaction is possible from the log messages. The retraining with the annotations was not done in this research, due to limited time for data collection.

By integrating the model into an interface, the radiologist can use, and also by leveraging the interface for collecting annotated data, we can increase the efficiency of the annotation process. This approach effectively addresses our research questions and offers a practical solution for enhancing both the accuracy and efficiency of radiology reporting.

6.2 Limitations

In this research, we proposed an approach of using Conditional Random Fields and regular expression rules to ensure completeness and compliance with prostate screening radiology reports. After the development of the model, we integrated it into the workflow of radiologists using an interface to facilitate its practical application. One of the key challenges was having a limited annotated dataset, which led us to use the same interface to collect user interactions during the reporting process itself. However, there were several limitations in our study:

1. **Adaptation to Other Domains**

   The model was designed on prostate screening reports, which limits its application to other radiology reports. Future work would be needed to test if the model can be adapted to other domains of radiology by training on the specific data.

2. **Limited Evaluation of Interactive Interface**

   Although the evaluation was performed, and observations were obtained, still another set of iterative evaluations after incorporating the suggestions and findings from the previous evaluation, would give more insights about the interface usability.

3. **Integration of the Interface in Real-Time**

   We were unable to integrate the interactive interface into the hospital system fully. We use the non-interactive interface that shows the model results into the system, but to know the performance of the interactive one in real-time instead of at the end of reporting is not still done. Thus the potential impact of such a feature on both reporting efficiency and model accuracy in the long run remains unexplored.

4. **Model Performance on Varying Medical Terminology**

   The reporting contents can change over time. Small changes were noticed in the training dataset and the dataset used for the evaluation of the interface itself. Using the annotation from the user interaction can help in capturing the new changes, but still, an extensive study would be required to test how well the model adapts to the new changes.

5. **Radiologist Workload and Acceptance**

   We evaluated the interface with the radiologists and analyzed their experience and feedback. But this was done with two radiologists and in a short period. From this evaluation, we got positive feedback that using the interface will reduce the burden of annotation and minimize the burden on radiologists, its long-term impact on their workload has not been explored. As mentioned before, this can be only tested by using the interface along with their reporting over time, to understand whether radiologists can adapt to it and find it a helpful aid.

6. **Limited Exploration of Model Retraining**

   The user interaction data collected were relatively small due to time constraints, limiting the ability to demonstrate significant improvements in model performance with the additional data. Also, factors like how frequently the models should be retrained have to be explored in future scope.

Additionally, while the overfitting issue with language models was partially attributed to limited annotated data, further exploration into this issue was beyond the scope of this thesis. Our primary focus was on designing an effective annotation interface to address the problem, though other potential solutions remain to be explored.

6.3 Recommendations

Based on the findings of this research, the following recommendations are proposed for future work:

1. **Adapting to Other Radiology Reporting**

   While this research concentrated on prostate screening reports, the approach can be extended to other radiology reporting, by adjusting the model and Regular Expression (RE) rules accordingly. This involves training domain-specific datasets to capture relevant features and improve extraction accuracy.

2. **Large Language Models**

   In this research, the use of the Language model was not a success. The limited dataset may be a reason. As more data is collected through the interface, language models can be revisited and retrained with the annotated data. With sufficient data, these models may exhibit better performance and adaptability, making them a promising area for future exploration.

3. **Integrating into the Existing System**

   The current interface can be developed to integrate with the hospital systems to support real-time checking of the reports and collection of annotations. An iterative evaluation of the interface has to be conducted to get more insights into the usability of the interface.

4. **User Training for Interactive Interface**

   Providing the radiologists with training in how to effectively use the interface could improve user adoption. Demonstrating how the interface can save time and ensure report quality would encourage more radiologists to use the interface.

5. **Speech-Based Interactions**

   Exploring the implementation of speech-based interaction instead of clicks can be interesting, as the radiology reporting is speech-based. Thus speech-based interactions will be easier for the user. This involves exploring how speech commands can be used to navigate and interact with the reports.

Bibliography

[1] M. P. Hartung, I. C. Bickle, F. Gaillard, and J. P. Kanne, “How to create a great radiology report,” *Radiographics*, vol. 40, no. 6, pp. 1658–1670, 10 2020.

[2] A. P. Brady, “Radiology reporting—from Hemingway to HAL?” pp. 237–246, 4 2018.

[3] L. Zhang, X. Wen, J. W. Li, X. Jiang, X. F. Yang, and M. Li, “Diagnostic error and bias in the department of radiology: a pictorial essay,” 12 2023.

[4] Michael Walter, “Incomplete US radiology reports lead to confusion, unnecessary biopsies.” [Online]. Available: https://radiologybusiness.com/topics/health-it/enterprise-imaging/imaging-informatics/incomplete-radiology-reports-lead-confusion

[5] Kasalak, H. Alnahwi, R. Toxopeus, J. P. Pennings, D. Yakar, and T. C. Kwee, “Work overload and diagnostic errors in radiology,” *European Journal of Radiology*, vol. 167, 10 2023.

[6] A. P. Brady, “Error and discrepancy in radiology: inevitable or avoidable?” pp. 171–182, 2 2017.

[7] M. Pahadia, S. Khurana, H. Geha, and S. T. Deahl, “Radiology report writing skills: A linguistic and technical guide for early-career oral and maxillofacial radiologists,” *Imaging Science in Dentistry*, vol. 50, no. 3, pp. 269–272, 9 2020.

[8] N. werkgroep prostaat MRI., *Kwaliteitsdocument Prostaat MRI: protocol en verslaglegging*. Nederlandse Vereniging voor Radiologie, 2023. [Online]. Available: https://radiologen.nl/system/files/bestanden/documenten/kwaliteitsdocument_prostaat_mri_protocol_en_verslaglegging_av_nov_2023_def.pdf

[9] N. Linna and C. E. Kahn, “Applications of natural language processing in radiology: A systematic review,” 7 2022.

[10] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, "Natural language processing technologies in radiology research and clinical applications," *Radiographics*, vol. 36, no. 1, pp. 176–191, 1 2016.

[11] S. Pathak, J. Van Rossen, O. Vijlbrief, J. Geerdink, C. Seifert, and M. Van Keulen, "Post-Structuring Radiology Reports of Breast Cancer Patients for Clinical Quality Assurance," Tech. Rep.

[12] F. H. Chokshi, A. E. Flanders, L. M. Prevedello, and C. P. Langlotz, "Fostering a healthy ai ecosystem for radiology: Conclusions of the 2018 rsna summit on ai in radiology," *Radiology: Artificial Intelligence*, vol. 1, no. 2, 3 2019.

[13] B. Allen, S. E. Seltzer, C. P. Langlotz, K. P. Dreyer, R. M. Summers, N. Petrick, D. Marinac-Dabic, M. Cruz, T. K. Alkasab, R. J. Hanisch, W. J. Nilsen, J. Burleson, K. Lyman, and K. Kandarpa, "A Road Map for Translational Research on Artificial Intelligence in Medical Imaging: From the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop," *Journal of the American College of Radiology*, vol. 16, no. 9, pp. 1179–1189, 9 2019.

[14] P. López-Úbeda, T. Martín-Noguerol, K. Juluru, and A. Luna, "Natural Language Processing in Radiology: Update on Clinical Applications," *Journal of the American College of Radiology*, vol. 19, no. 11, pp. 1271–1285, 11 2022.

[15] E. Nguyen, D. Theodorakopoulos, S. Pathak, J. Geerdink, O. Vijlbrief, M. Van Keulen, and C. Seifert, "A Hybrid Text Classification and Language Generation Model for Automated Summarization of Dutch Breast Cancer Radiology Reports," Tech. Rep. [Online]. Available: https://github.com/daphne12345/SummarizationRadiologyReports

[16] L. F. Donnelly, R. Grzeszczuk, and C. V. Guimaraes, "Use of Natural Language Processing (NLP) in Evaluation of Radiology Reports: An Update on Applications and Technology Advances," *Seminars in Ultrasound, CT and MRI*, vol. 43, no. 2, pp. 176–181, 4 2022.

[17] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque, "Named Entity Recognition in Electronic Health Records: A Methodological Review," pp. 286–300, 10 2023.

[18] S. Bozkurt, E. Alkim, I. Banerjee, and D. L. Rubin, "Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 544–553, 8 2019.

[19] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, “BioBERT based named entity recognition in electronic medical record,” in *Proceedings - 10th International Conference on Information Technology in Medicine and Education, ITME 2019*. Institute of Electrical and Electronics Engineers Inc., 8 2019, pp. 49–52.

[20] S. Verkijk, P. Vossen, and P. T. J. M. V. Nl, “MedRoBERTa.nl: A Language Model for Dutch Electronic Health Records,” Tech. Rep., 2021. [Online]. Available: https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health

[21] M. T. Rietberg, V. B. Nguyen, J. Geerdink, O. Vijlbrief, and C. Seifert, “Accurate and Reliable Classification of Unstructured Reports on Their Diagnostic Goal Using BERT Models,” *Diagnostics*, vol. 13, no. 7, 4 2023.

[22] B. Romera-Paredes and P. H. S. Torr, “An embarrassingly simple approach to zero-shot learning,” Tech. Rep.

[23] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical Networks for Few-shot Learning,” 3 2017. [Online]. Available: http://arxiv.org/abs/1703.05175

[24] S. Arora, A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré, “Ask Me Anything: A simple strategy for prompting language models,” 10 2022. [Online]. Available: http://arxiv.org/abs/2210.02441

[25] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen, “Making Large Language Models Better Reasoners with Step-Aware Verifier,” 6 2022. [Online]. Available: http://arxiv.org/abs/2206.02336

[26] M. Rossini Paolo Torroni and E. Cabrio Serena Villata, “PROMPTING TECHNIQUES FOR NATURAL LANGUAGE GENERATION IN THE MEDICAL DOMAIN CANDIDATE SUPERVISOR,” Tech. Rep.

[27] Y. Wang, Y. Wang, Z. Peng, F. Zhang, L. Zhou, and F. Yang, “Medical text classification based on the discriminative pre-training model and prompt-tuning,” *Digital Health*, vol. 9, 1 2023.

[28] Y. Lu, X. Zhao, and J. Wang, “Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text,” Tech. Rep., 2023.

[29] T. Wu, D. Yang, and S. Santy, “Designing, Learning from, and Evaluating Human-AI Interactions,” Tech. Rep. [Online]. Available: https://xai-hcee.github.io/

[30] Z. J. Wang, D. Choi, S. Xu, and D. Yang, "Putting Humans in the Natural Language Processing Loop: A Survey," 3 2021. [Online]. Available: http://arxiv.org/abs/2103.04044

[31] P. C. Lo and E. P. Lim, "Interactive Entity Linking Using Entity-Word Representations," in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, 7 2020, pp. 1801–1804.

[32] P. P. Gaurav Trivedi, "An interactive tool for natural language processing on clinical text," https://arxiv.org/pdf/1707.01890, 2017, [Accessed 17-09-2024].

[33] "Inception Platform." [Online]. Available: https://inception-project.github.io/

[34] A. B. T. C. G. v. N. Wietse de Vries, Andreas van Cranenburgh and M. Nissim, "Bertje: A dutch bert model," https://arxiv.org/pdf/1912.09582, 2019, [Accessed 30-07-2024].

[35] J. G. O. V. a. S. Max Tigo Rietberg, Van Bach Nguyen, "Accurate and Reliable Classification of Unstructured Reports on Their Diagnostic Goal Using BERT Models — doi.org," https://doi.org/10.3390/diagnostics13071251, 2023, [Accessed 30-07-2024].

[36] A. S. Albert Q. Jiang, "Mistral 7b," https://arxiv.org/pdf/2310.06825, 2023, [Accessed 17-09-2024].

[37] A. M. Charles Sutton, "An introduction to conditional random fields," https://arxiv.org/pdf/1011.4088, 2010, [Accessed 17-09-2024].

[38] "Figma." [Online]. Available: https://www.figma.com/

[39] "Welcome to Flask &x2014; Flask Documentation (3.0.x) — flask.palletsprojects.com," https://flask.palletsprojects.com/en/3.0.x/, [Accessed 15-08-2024].