A Hybrid Text Classification and Language Generation Model for Automated Summarization of Dutch Breast Cancer Radiology Reports

Elisa Nguyen*‡, Daphne Theodorakopoulos*‡, Shreyasi Pathak*, Jeroen Geerdink†, Onno Vijlbrief†, Maurice van Keulen* and Christin Seifert*§
* *Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands*
† *Hospital Group Twente, Hengelo, The Netherlands*
§ *Faculty of Medicine, University of Duisburg-Essen, Essen, Germany*
Email: <t.q.e.nguyen,d.theodorakopoulos>@student.utwente.nl,
<s.pathak,m.vankeulen,c.seifert>@utwente.nl,
<j.geerdink,o.vijlbrief>@zgt.nl
‡ *Authors contributed equally to this work.*

2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) | 978-1-7281-4144-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/COGMI50398.2020.00019

***Abstract*—Breast cancer diagnosis is based on radiology reports describing observations made from medical imagery, such as X-rays obtained during mammography. The reports are written by radiologists and contain a conclusion summarizing the observations. Manually summarizing the reports is time-consuming and leads to high text variability. This paper investigates the automated summarization of Dutch radiology reports. We propose a hybrid model consisting of a language model (encoder-decoder with attention) and a separate BI-RADS score classifier. The summarization model achieved a ROUGE-L F1 score of 51.5% on the Dutch reports, which is comparable to results in other languages and other domains. For the BI-RADS classification, the language model (accuracy 79.1%) was outperformed by a separate classifier (accuracy 83.3%), leading us to propose a hybrid approach for radiology report summarization. Our qualitative evaluation with experts found the generated conclusions to be comprehensible and to cover mostly relevant content, and the main focus for improvement should be their factual correctness. While the current model is not accurate enough to be employed in clinical practice, our results indicate that hybrid models might be a worthwhile direction for future research.**

***Keywords*-Abstractive Summarization, Radiology Reports, Breast Cancer, Deep Learning, Encoder-Decoder, Attention Mechanism**

I. Introduction

Mammography is one of the diagnostic tests performed for diagnosing breast cancer and findings of these mammography images along with findings from some other diagnostic tests like ultrasound and magnetic resonance imaging (MRI) are documented by radiologists in radiology reports. The reports need to be clear and consistent so that the findings can be easily understood by other physicians. The Breast Imaging Reporting & Data System (BI-RADS) is used as a standard in breast cancer reporting [1] specifying the structure of the reports. Reports adhering to this standard consist of clinical data, findings from the examinations as well as a conclusion including a BI-RADS score (ranging from 0 to 6, where 6 is the most severe malignancy). An example report is shown in Figure 1. Radiologists write (or dictate) these reports in free text, leading to variability of the structure and writing quality of the reports. Besides, the findings are often written during the examination, and it is time-consuming to write a conclusion as patient data needs to be consulted again. Therefore, a system for summarization of the findings in the form of an automatic conclusion can speed up the diagnostic process and contribute to human error reduction and consistency of reports. Further, a system that has learned the relationship between report content and conclusion could be used for quality control and consistency checks of radiology reporting in a hospital.

Original	**English Translation**
Input sequence: *medische gegevens:* via SVOB, microcalcificaties R lateraal boven *verslag:* matig beoordeelbaar dens klierweefsel beiderzijds, microcalcificaties laterale bovenkwadrant rechtermamma overgang laterale onderkwadrant diameter 2,3 cm, stellate laesies, echografisch onderzoek axilla rechts laat geen pathologische lymfomen zien **Ground Truth Conclusion:** microcalcificaties in het laterale bovenkwadrant van de rechtermamma, birads-classificatie-iv, geen pathologische lymfomen in de axilla **Generated Conclusion:** birads iv laesie in het laterale bovenkwadrant van de rechtermamma waarvoor advies stereotactische biopsie	**Input sequence:** *clinical data:* via SVOB, microcalcifications R lateral upper *findings:* The breasts are heterogeneously dense on both sides, microcalcifications in the lateral upper quadrant of the right breast at the junction of the lateral lower quadrant. Diameter 2.3 cm, stellate lesions. Ultrasound of right axilla shows no pathological lymph nodes. **Ground Truth Conclusion:** Microcalcifications in the lateral upper quadrant of the right breast. BIRADS classification IV. No pathological lymph nodes in the axilla. **Generated Conclusion:** BIRADS IV lesion in the lateral upper quadrant of the right breast requiring stereotactic biopsy.

Figure 1. Example of a report containing the findings, the original and generated conclusion of the EDA+BI-RADS model. Dutch on the left, English translation on the right (Translated by a radiologist).

Although the BI-RADS standard [1] asks for a clear structure, reports found in practice are partly unstructured, do not consist of full sentences, and include typing errors [2]. Therefore, it is unclear to what extent existing automatic summarization methods for general texts [3], [4] can result in high-quality summaries on unstructured medical reports.

978-1-7281-4144-2/20/$31.00 ©2020 IEEE
DOI 10.1109/CogMI50398.2020.00019

MacAvaney et al. [5] applied a pointer-generator network (PGN) [6] to generate the impression (summary) of English radiology notes from a variety of imaging modalities. In contrast to their summaries, our summaries should follow the BI-RADS standard, which requires the summary to consist of a classification (the BI-RADS score) and a concluding sentence. In this paper, we compare a state-of-the-art summarization method for generating the conclusion including the BI-RADS score, and a hybrid model, where a text classifier is used to predict the BI-RADS score separately, which is then integrated into the summary produced by the state-of-the-art summarization method.

More specifically, we use an encoder-decoder model with attention (EDA) trained on Dutch radiology reports. This model is compared to a baseline model without attention (ED). Additionally, we investigate the prediction of the BI-RADS score separately using classical text classification in TF-IDF[1] vector space. We evaluate the summaries w.r.t. ROUGE scores [7], the accuracy of the BI-RADS score in the summary, and perform an expert evaluation to judge the correctness, relevancy, and comprehensibility of the generated reports. We found the hybrid model, which inserts the BI-RADS score classification results in the abstractive summaries, to outperform the pure summarization models by 5% in the accuracy of BI-RADS score prediction. As the BI-RADS score is the most important clue for subsequent treatment, our results indicate that a combination of multiple models (classification and text summarization) is a worthwhile direction for future research. The source code is available on GitHub[2].

The remainder of the paper is structured as follows. Section II presents related work. Section III describes the algorithms and data sets. Results are shown in Section IV and discussed in Section V.

II. Related Work

In this section, we discuss some existing works on BI-RADS classification and automatic text summarization.

A. BI-RADS Score Classification

The majority of existing methods for the classification of BI-RADS scores using natural language processing (NLP) are rule-based and extract specific features. Sippo et al. [8] presented an NLP-tool for BI-RADS score classification. It determines the BI-RADS scores through regular expressions and string matching of selected parts of the report after some preprocessing of the text. Their study involved training and testing their model on 1165 instances of data from a breast imaging center in the United States and achieved an overall F1 score of 98%.

Castro et al. [9] enhanced this approach by extracting certain features such as imaging study type and laterality of the breast relevant for BI-RADS score classification. Their best model using rules from partial decision trees was able to achieve an overall F1 score of 91% trained and tested on a larger dataset (2159 instances of data) of 18 hospitals in Pittsburgh.

Banerjee et al. [10] proposed a semi-supervised NLP pipeline for retrieving the BI-RADS score from mammography reports. They used semantic dictionary mapping that assigns the words in the reports to key terms. These should capture the true semantics of the report and therefore facilitate better information extraction while keeping a low dimensionality of information representation. They used a logistic regression classifier, which achieved an overall F1 score of 89% on the classification task. Instead of extracting specific features as in these works, we are using a generic TF-IDF approach to determine important words as features.

B. Automatic Text Summarization

Text summarization is categorized into abstractive and extractive methods. Our work uses abstractive summarization, which means that novel sentences are built from the vocabulary. This method stands in contrast to extractive summarization, which copies the most important sentences from the input to generate a summary [11], [12].

There are different approaches to solve the task of abstractive summarization. Most of them are based on deep learning using sequence-to-sequence (seq2seq) models, that are composed of an encoder and decoder. The encoder maps the input to a context vector and the decoder generates the summarized target sequence word-by-word. One of the first works using seq2seq models for natural language generation was done by Sutskever, Vinyals and Le [13]. Similar to their work, we also use multilayered Long Short-Term Memory (LSTM) in the encoder and decoder in our work. The method is often further enhanced by using an attention mechanism which was first introduced in [14]. The attention mechanism considers the hidden state of the encoder and the decoder. The learned weights tell the decoder the parts of the input sequence to pay attention to produce the next word The first studies applying seq2seq in combination with attention to the task of summarization were Rush, Chopra and Weston [15], and Nallapati et al. [4]. The dominant sequence transduction models are based on the deep encoder-decoder structure with attention [16]. The model used in our work also uses a seq2seq model with attention.

In the biomedical domain, work has been focused on enhancing current state-of-the-art (SOTA) models for abstractive summarization with domain-specific knowledge. Examples are summarization of biomedical publication abstracts and electronic health reports [17], [18]. A recent study [5] extended the PGN model with domain-specific ontological information from existing medical ontologies such as RadLex. The ontology-linked entities in the report were provided as a separate context vector to the decoding

[1]TF: Term Frequency; IDF: Inverse Document Frequency

[2]https://github.com/daphne12345/SummarizationRadiologyReports

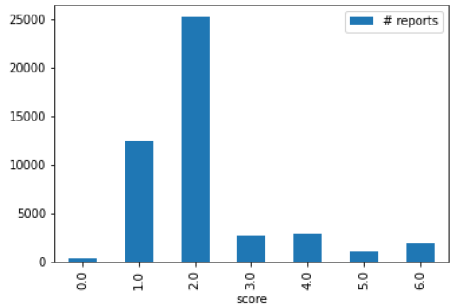

Figure 2. Number of reports per BI-RADS score in the entire dataset

process. By including domain-specific knowledge, it was found that the summaries from the extended models are statistically better than the general SOTA models on radiology corpora, achieving a ROUGE-L score of 37.02%. This study illustrates the potential of abstractive summarization in radiology to which our work is contributing.

In conclusion, there is not much work focusing on summarizing Dutch medical reports and our work is among the first to use a model similar to the state-of-the-art abstractive summarization model [14] on Dutch radiology reports and show some interesting insights in this direction.

III. Data and Approach

In this section, we describe our dataset and method for automatic summarization of the radiology reports.

A. Data and Preprocessing

Our **dataset** comprises 47,158 breast cancer radiology reports from the Ziekenhuis Groep Twente (ZGT), a hospital in Hengelo, Netherlands, recorded between 2012 and 2018. The reports are in Dutch and are written in free text. The reports include *clinical data* (indication for this diagnostic study including patient's medical history), *findings* (clinical findings from the diagnostic images - mammography, ultrasound and MRI) and *conclusion* (Final assessment including a BI-RADS score) (see Figure 1). The clinical data and findings are treated as the input sequence in the frame of this work. This information usually indicates the breast cancer severity (BI-RADS score) which is relevant for the conclusion. The class (BI-RADS score) distribution in our dataset shows that BI-RADS 2 is the majority class and BI-RADS 0 is the minority class (cf. Figure 2).

For **preprocessing** the data, first, stop words are removed. As we are dealing with a Dutch corpus, the Dutch stop words from NLTK [19] were used for this task. A tokenizer is used to create the vocabulary. For the conclusions, start and end tokens are added. Secondly, each word in the vocabulary is represented by an index. So, both findings and conclusions are sequences of numbers. The maximum length of the findings representing the input sequence is limited to 100 due to the availability of computational resources. The maximum length of the conclusion is set to 32, to ensure a reasonable length. This number was determined based on the ratio of the median lengths of findings and conclusion (46:12) $\times$ 100, plus a few words as a buffer. If the findings and conclusions are shorter than 100 and 32 respectively, they are padded with zeros. This represents the required fixed length of the context vector given from encoder to decoder.

For the BI-RADS classification, further preprocessing is needed. The BI-RADS score needed to be extracted from the given conclusions as they were not given in the dataset as a separate attribute. The dataset contains data from different radiologists which means that there is no common way of reporting the BI-RADS score in the different reports. Different number formats (i.e. 2, ii, twee) have been used. Also, sometimes a word is inserted in between (e.g., "BI-RADS rechts 2") or it has been indicated differently (e.g., "BI-RADS classificatie 2", "BI-RADS-ii"). We constructed a set of rules for extracting all variants of the BI-RADS scores.

B. Model

In this subsection, we will describe our 3 models - i) text summarization model (we compared a baseline encoder-decoder model with an encoder-decoder-attention model), ii) BI-RADS classification model, and iii) hybrid model (text summarization + BI-RADS classification). The text summarization models were used for generating the conclusion of the radiology reports from the clinical data and findings in the reports. To get a more accurate BI-RADS score (to be included in the conclusion part of the report), a separate BI-RADS classifier was trained. To have a combined model that contains an accurate BI-RADS score and generated conclusion, a hybrid model was created combining the power of the above models. An overview picture of our model can be found in Figure 3.

1) Encoder-Decoder (ED) (Baseline model): In abstractive text summarization, seq2seq models [13] are often used mapping an input sequence to an output sequence of different lengths using an encoder and a decoder. The input sequence is passed through a word embedding layer which maps the numerical input received from preprocessing to embeddings. The embedding is given to the LSTM-based encoder. The LSTM units gather information about the elements in the input sequence and propagate the information forward in the sequence. The output of the encoder is the context vector containing the encoder hidden state with a fixed length, which serves as the initial hidden state of the decoder, which is a representation of the target sequence. The decoder also contains several LSTM units. The decoder iterates through the context vector and predicts the next word (the one with the highest probability) given the previous word, and starting with a `_START_` token. The sequence

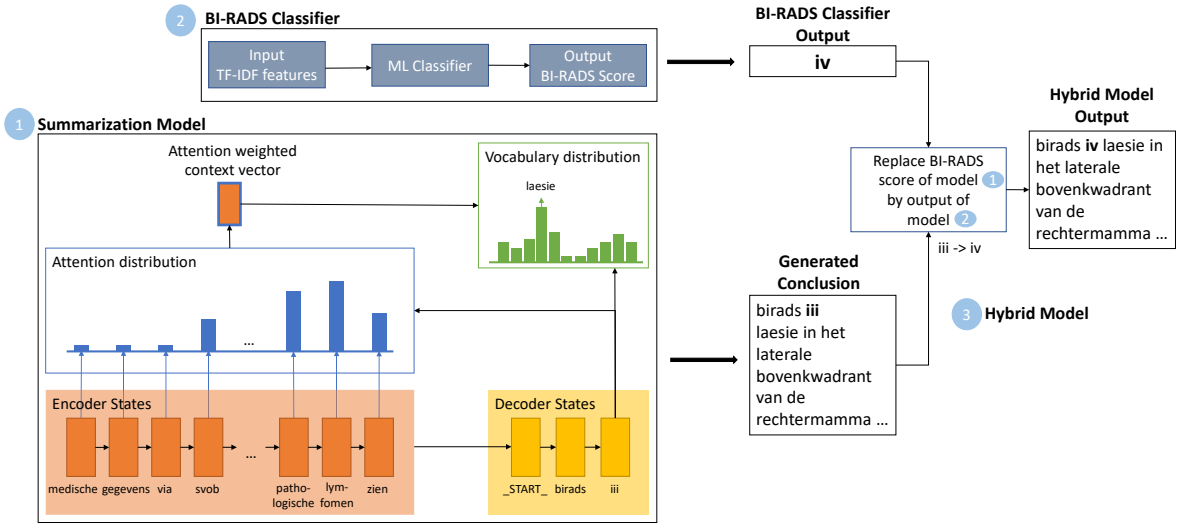

Figure 3. Conclusion generation with hybrid model - combination of summarization model (Encoder-Decoder-Attention model) (Own diagram based on [6]) and BI-RADS classifier.

ends once the end token is predicted or the maximum decoding length of 32 has been reached. This is done by calculating the probability using $softmax$ over all the words in the vocabulary at each decoding step. The word in the vocabulary with the largest probability is chosen as the next decoded token. In Figure 3, the baseline ED model is represented by the orange (encoder), yellow (decoder) and green (output vocabulary distribution) parts present in the part labeled 1.

2) Encoder-Decoder-Attention Model (EDA): The EDA model differs from the ED model by the addition of an attention layer [14], which allows the decoder to have access to all hidden states of the encoder at each decoding step. In the ED model, the context vector passed from the encoder to the decoder has a fixed length and represents the result of the last encoding unit. Thus, all previous hidden encoder states are discarded. This leads to a disadvantage when using seq2seq models, as longer sequences tend to be squeezed into this fixed-length context vector and information can be lost. The mechanism of attention [14] is intended to solve this issue. It makes use of these states during the decoding process. The hidden encoder states are attended to, depending on the current state of the decoder. This means that certain words of the input sequence are considered in addition. The model assigns high attention scores to those words of the input sequence that are relevant to the current decoder step. In Figure 3, the part with label 1 shows the EDA summarization model with an attention layer (indicated in blue) added to the architecture of the ED model.

We illustrate the approach with an example summarization step in Figure 3 using the example report of Figure 1. Note that the words "pathologische" (pathological) and "lymfomen" (lymphoma) in the report receive higher attention scores at the decoding step. In concatenation with the context vector resulting from the encoder, this attention-weighted context vector is fed to the next decoding unit, which chooses "laesie" (lesion) as the next word from the vocabulary. The blue part of Figure 3 shows the attention of the example.

The **architecture** of the **encoder** in both ED and EDA models consists of an input layer and three LSTM layers to capture complex input whereas the **decoder** consists of an LSTM layer, an attention layer, and a dense output layer. All LSTM layers follow the default setting of $tanh$ as activation function and $sigmoid$ as the input/output/forget gate activation function. The implementation of the neural network was mostly done with the library Keras [20]. The attention layer from [14] was implemented.

3) BI-RADS Classification: The summarization model is focused on the generation of text. A common problem with language generation models is that they often do not reproduce facts correctly. Therefore, a second model was used which solely focuses on the correct prediction of the most important fact: the BI-RADS score.

In essence, predicting the BI-RADS score from the clinical data and findings is a text classification problem. We use the common 'bag of words' approach where we use TF-IDF [21] scores as the values for the word features. The TF-IDF scores inform the classifier about the *distinctiveness* of words. If a term (word) occurs in many documents (clinical data + findings), it is not very distinctive for a class of documents, thus it has a low TF-IDF. In contrast, if a word only occurs in a few findings with a high frequency, it is likely to be distinctive for that class and is quite informative,

thus it has a high TF-IDF. The TF-IDF matrix is generated from clinical data and findings of a report and is passed as an input feature to a traditional machine learning (ML) classifier to predict the BI-RADS score, which can take any categorical value in the range of 0-6. Figure 3 contains the BI-RADS classifier (shown with label 2).

4) EDA+BI-RADS (Hybrid Model): The hybrid model first uses the input sequence of clinical data and findings to generate the conclusion using the best summarization model. Then, the BI-RADS classifier predicts the BI-RADS score based on the clinical data and the findings using the classification pipeline. Finally, the BI-RADS score in the generated conclusion of the summarization model is textually replaced by the prediction of the BI-RADS classifier. Figure 3 shows how the hybrid model is formed (shown with label 3). The output of the model can be seen in the 'Hybrid Model Output' on the right in the figure.[3]

C. Experimental Setup

The hybrid summarization and BI-RADS classifier model is experimentally compared with the non-hybrid EDA and ED baseline models. For training, hyper-parameter tuning, and testing, the dataset has been divided into a train (70%), validation (10%), and test (20%) set.

1) Summarization models: By summarization models, we refer to both ED and EDA models. The model has been fitted using early stopping to prevent over-fitting. The patience parameter was set to 20 epochs to avoid stopping the training prematurely at a local optimum. This means that only after 20 epochs of no improvement measured by loss on the validation set, the training will be stopped. After early stopping, the weights of the best epoch are restored. The sparse-categorical-cross-entropy loss function as well as the RMSprop optimizer have been applied. Sparse-categorical-cross-entropy is appropriate as a loss function because the problem at hand is a multi-class problem. The RMSprop optimizer was chosen as it is known to deal well with mini-batches during training, which is the case.

This configuration of the model has two hyper-parameters that need tuning: *Latent dimension*, which is the number of hidden units in an LSTM layer, and the *batch size* during training. For latent dimensions, we used different values roughly around the size of the input and output sequence, which are 100 and 32 respectively. Hence, the following values were used in tuning of the latent dimension: 60, 80, 100, 120. For the batch size, we decided on the values of 64, 128, 256, and 512 for hyper-parameter tuning. The batch size should not be too small so that the model has enough data to find patterns. It cannot be too large either because the training time increases drastically. To cover the entire parameter space, a grid search was performed.

[3]The summarization model always generates a BI-RADS score in its conclusion.

2) BI-RADS Classification: We compared different multi-class classifiers for BI-RADS classification: Support Vector Machine (SVM), Logistic Regression, Ridge Classifier, Gradient Boosted Trees, Random Forest, K-Nearest Neighbour (KNN) and Multinomial Naive Bayes. We used the python implementations in Scikit-Learn [22] and Xgboost [23].

3) Evaluation Metrics: We evaluated the generated conclusions quantitatively and qualitatively.

Quantitative: To assess the performance of the ED and EDA models, ROUGE scores [7] are used. ROUGE scores are measures used to evaluate the quality of texts to an ideal reference text. A ROUGE-n score counts the numbers of overlaps of respective n-grams between the reference and the text at hand. High scores indicate a high overlap between the prediction and reference and are therefore desirable. In the frame of this work, the measures of precision, recall, and F1 of ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common subsequence) are determined between the reference conclusion from the test set and the generated conclusion by using the py-rouge library [24]. Moreover, the generated conclusion is anticipated to provide a BI-RADS score as well. Therefore, this score is extracted and the accuracy of the BI-RADS score classification from the summarization is reported as well. All ROUGE scores and BI-RADS score prediction accuracy on the test set for ED, EDA and BI-RADS classifiers are reported for 95% confidence interval calculated using $1.96\sqrt{\frac{v(1-v)}{N}}$, where v is the score or the accuracy value, N is the number of reports in the test set and 1.96 is the z-score for 95% confidence.

The fine-tuning of the hyper-parameters is done on the validation set. The three best performing combinations in terms of ROUGE-L F1 are further evaluated on the test set. For these models, the ROUGE-L F1 score and BI-RADS accuracy are calculated. This is done to determine the best hyperparameter combination of the summarization model and validate the need for the separate BI-RADS classifier. We chose the F1 score as it combines precision and recall.

For the evaluation of the BI-RADS classifier, each of the pipelines with the different classifiers was trained on the training set and evaluated on the validation and test set using accuracy as a metric.

Qualitative: In addition to these metrics, a qualitative evaluation is done with two radiologists of the ZGT hospital in Hengelo, Netherlands. The objective of this evaluation is to get an impression on whether the content is *correct* (factual correctness of the information), *relevant* (medical relevance of the content for the doctors) and *comprehensible* (makes sense syntactically and semantically). For this, a random sample of five pairs of original findings and its generated conclusions are sent to the radiologists. Both the radiologists discussed and rated the generated conclusions together in terms of the 3 aforementioned criteria. Furthermore, they were asked to give free-text comments about each

Table I
EDA PERFORMANCE FOR DIFFERENT BATCH SIZES AND LATENT DIMENSIONS ON THE TEST SET.

Batch size	Latent dimension	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
128	120	0.540	0.388	**0.515**
64	100	0.531	0.381	0.508
64	120	0.526	0.377	0.503

conclusion.

IV. RESULTS

Table IV in Appendix A shows the performance of the different models, resulting from the hyperparameter tuning of the EDA model on the validation set. On evaluating the 3 best performing combinations of hyperparameters on the test set, we found that batch size of 128 and latent dimension of 120 achieved the highest ROUGE-L F1 (cf. Table I). Our ED and EDA models were trained on the best hyperparameters and the resulting ROUGE scores are reported in Table II. Our EDA model outperformed the baseline ED model by around 0.7%, achieving a ROUGE-1 F1 score of 0.54, a ROUGE-2 F1 score of 0.388, and a ROUGE-L F1 score of 0.515. However, the ED model achieved a 1% higher BI-RADS score accuracy compared to the EDA model.

For the classification of the BI-RADS score, seven different supervised learning methods were inserted in the pipeline and then compared on the validation set. The accuracy of all the classifiers on the validation set is shown in Table III. The three best performing models (SVM, Logistic Regression and Ridge classifier) were also compared on the test set and the accuracy can be found in Table III. SVM classifier was found to be the best performing BI-RADS classifier. The BI-RADS classification accuracies from the ED and EDA models are stated in Table II for comparison and it can be seen that SVM outperforms the ED model by 4%.

We have also shown a generated conclusion from a fictive report using our hybrid model in Figure 1 (shown both the original Dutch report along with its English translation). As can be seen, there are many common terms between our generated conclusion and the ground truth conclusion, e.g. 'birads', 'iv', 'lateral', 'upper', 'quadrant', 'of', 'the', 'right', 'breast'. Figure 3 shows that the EDA model generated a wrong BI-RADS score of *iii* (three) in the generated conclusion and the BI-RADS classifier predicted the BI-RADS score correctly as *iv* (four). Therefore, in our hybrid model output, replacing the BI-RADS score of *iii* from the EDA model with the BI-RADS score of *iv* from the BI-RADS classifier results in a correct BI-RADS score.

The results of the qualitative evaluation with the hospital can be found in Figure 4. The five example reports had a correctness of 40%, relevancy of 75%, and were 85% comprehensible. The comments from the radiologists indicated

Table II
COMPARISON OF THE ED AND EDA MODELS FOR THE SUMMARIZATION TASK ON THE TEST SET BASED ON ROUGE SCORE AND BI-RADS ACCURACY REPORTED AT 95% CONFIDENCE INTERVAL.

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BI-RADS Accuracy
ED	0.530±0.01	0.383±0.01	0.508±0.01	**0.791±0.01**
EDA	**0.540±0.01**	**0.388±0.01**	**0.515±0.01**	0.784±0.01

Table III
COMPARISON OF DIFFERENT BI-RADS CLASSIFIERS ON VALIDATION AND TEST SET (TEST SET VALUES REPORTED AT 95% CONFIDENCE INTERVAL). THE BASELINE IS AN ARTIFICIAL CLASSIFIER THAT ALWAYS PREDICTS THE MAJORITY CLASS (BI-RADS SCORE 2).

Model	Accuracy	
	Validation	Test
SVM	**0.837**	**0.833±0.01**
Logistic Regression	0.817	0.816±0.01
Ridge Classifier	0.797	0.795±0.01
Gradient Boosted Trees	0.780	0.779±0.01
Random Forest	0.773	0.768±0.01
KNN	0.696	0.695±0.01
Multinomial Naive Bayes	0.670	0.675±0.01
Baseline Majority classifier	0.523	0.542

the problems in the generated conclusions, e.g. wrong breast side and absence of the word "geen" ("no" in English) from the findings. This leads to some of the conclusions meaning the opposite of the intended sense.

V. DISCUSSION

A. Interpretation of results

In this work, we automatically generated conclusions of Dutch radiology reports including a classification of the BI-RADS score. This was done by combining a summarization model (EDA) which generates the conclusion with a classification model. The hyperparameters of the EDA were tuned and the best performing model with a ROUGE-L F1 score of 0.515 has a batch size of 128 and a latent dimension of 120. Interestingly, other parameter combinations were performing better on the validation set. This could be due to overfitting during training. The ROUGE-L score is the most informative, as it combines the other two ROUGE scores. The score is quite high in comparison to similar works, such as [18] which also summarized medical data and had a ROUGE-L score of 0.347 for the same model as ours. Our significantly higher score on a similar task with the same model is probably due to shorter input texts of our dataset and less complexity. The SOTA summarization model "T5" had a ROUGE-L score of 0.407 on their best model [25]. Again, this difference can be explained with the given dataset.

The comparison of the final EDA model to the ED baseline model showed that the attention mechanism helps to generate a slightly better conclusion. The ROUGE-L F1

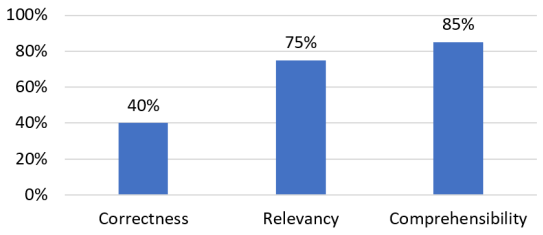

Figure 4. Hospital evaluation results. Average percentage of each generated conclusion being correct, relevant and comprehensible based on the ratings.

score of the simple encoder-decoder model scores only 0.7% lower. This rather small improvement could be explained by the length of the findings, which are at maximum 100 tokens. A larger increase in performance could be expected when longer sequences are passed as input. To quantify the uncertainty of the scores, we reported the scores at 95% confidence interval and we found the possibility of overlap in the scores from these two models. Therefore, no conclusion can be drawn on the question of whether or not attention can be beneficial in the given summarization task.

The accuracy of the EDA model in predicting the BI-RADS score was 78.4% as compared to the accuracy of 83.3% achieved by the separate BI-RADS classifier (SVM). Thus, a 4% improvement in accuracy indicates that a separate model for BI-RADS score classification is beneficial.

The ROUGE score only evaluates how well the generated conclusion text matches the human conclusion text, but it ignores several other important aspects of good conclusions such as correctness, relevance of the content, and understandability. The results from the hospital evaluation suggest that the generated conclusions are mostly factually incorrect (e.g. wrong breast side, wrong imaging protocol) but still cover relevant content and are comprehensible. We cannot report the agreement between the radiologists because they did the evaluation together and provided a combined score. The high scores in relevance and comprehensibility suggest that there is potential for an effective clinically usable automatic summarization approach if the approach can be improved on factual correctness.

While we have not used an explicit method to handle out-of-vocabulary (OOV) words, e.g. copy mechanism, our summarization model handles OOV words using the embedding layer present after the input sequence. Whenever the model encounters an OOV word, the embedding layer will find the word's closest meaningful representation in its vector space.

We adopted a neural text summarization model for the task of summarizing radiology reports in the Dutch language [14]. We refrained from using a more recent model, e.g. the T5 architecture [25], because of i) their computational intensity, and ii) the larger number of parameters that need to be fit. There are no pre-trained models available that are trained in medical Dutch, and our data set is comparably small. We used a standard architecture for our summarization task as our main objective was to test the potential of summarizing Dutch medical reports and whether a language model could learn to predict the BI-RADS score. We see the hybrid model with separate BI-RADs prediction as a solid baseline for future work.

B. Limitations

One of the complications of this work lies in the dataset. As the dataset is in Dutch, other pretrained models for abstractive summarization trained on English corpora could not be used. Therefore, a new model needed to be built. Additionally, the dataset encompasses real radiology reports made by humans. This means that the possibility of human errors in the dataset exists. The findings sometimes include another BI-RADS score than the reported score in the corresponding conclusion, which could mislead the model. This could be either a previous score from medical history or perhaps human error. Furthermore, the conclusions are also written by humans, so if there are errors, they are learned as well. Clinicians appear to have very diverse styles of reporting [26], so it is not surprising that the model was unable to find a standardized structure for the conclusions as well.

The labels for the BI-RADS score were extracted from the conclusions by looking for certain formats. We only checked a sample of BI-RADS score extractions manually, and cannot guarantee that our extracted labels are error-free. Moreover, in some of the conclusions, no score could be extracted because the phrasing was inconsistent and some conclusions contained human errors. Due to the complexity and variance of the underlying texts [26], neither a rule-based nor a machine-learning based approach is likely to be error-free.

The preprocessing eliminated stop words from the findings which usually improves the performance of the model because the model does not learn the context from irrelevant words. The comments of the qualitative evaluation mentioned that the word "geen" ("no" in English) is sometimes missing from the findings. Such words, should never be removed as they negate the meaning of subsequent content.

A limitation of the proposed summarization model is that it sometimes generates wrong facts: only two out of five example reports evaluated by the hospital reviewers were reported to contain correct content. Therefore, improving factual correctness is an important direction for future work.

Additionally, the generated text of the current method does not contain any grammar checks, so there is no guarantee for correct sentences. However, the original conclusions were also not written in full sentences.

Finally, the evaluation shows some limitations. Firstly, the used ROUGE metric is not a metric for the quality of the conclusion in terms of content. It assumes that the reference conclusions from the dataset are the target, whereas this

might not always be the case. Secondly, the qualitative evaluation with the radiologists from the hospital was done on a very small scale. Therefore, the results are only to be treated as an indicator for the conclusion quality in terms of comprehensibility, correctness of information, and relevance of content instead of a general result.

C. Future Work

To improve this work, future work can address the aforementioned limitations. It could look at applying the model on an English dataset, such as radiology reports within the MIMIC III database [27]. This would give a statement about the cross-lingual validity of the presented model. Also, pretrained models in English could be fine-tuned to this dataset to investigate transfer learning between languages.

Moreover, the summarization model can be improved by extending the data preprocessing. By further looking at the list of stop words, we might give the model some more valuable information that could improve performance. The introduction of a structure to the input report [2] and its influence on the model could be investigated. In addition, the model could be extended with more clinically relevant features (e.g., breast size) to ensure relevant information in the conclusion. These features could be identified in cooperation with medical staff with the required domain knowledge. A structure for the targeted conclusion could also be established together with medical staff so that a standardized way of reporting the conclusion can be automated. Additionally, the use of domain-specific ontologies can be investigated for feature extraction. Also, the conclusion could be checked for grammar after generation. Furthermore, the summarization model itself could be extended by the calculation of a generation probability to build a PGN [6]. To improve the accuracy of the BI-RADS score classifier, relevant features such as breast composition could be considered. Moreover, a neural network, solely used for the task of predicting the BI-RADS score, might achieve higher accuracy.

In our hybrid model, we directly replace the BI-RADS score of the EDA model with the BI-RADS score predicted by the classifier. Future work could investigate a tighter integration of the classifier and the language model, e.g., by employing multi-task learning strategies.

Future work could also improve on the evaluation methods used in this study. As a ground truth for the BI-RADS scores did not exist separately, we needed to extract it ourselves. This ground truth was used for evaluation, so future work would need to assess its validity. In general, this work shows that there is a need for a more informative evaluation metric than the current ROUGE metric which is widely used for summarization tasks. Furthermore, a larger scale qualitative evaluation could be done with radiologists, in order to understand the applicability of the generated conclusion better.

VI. Conclusion

In this work, a hybrid model (EDA+BI-RADS) was introduced for the automatic generation of conclusions including a BI-RADS score classification for medical findings of Dutch breast cancer reports. An encoder-decoder model with attention was combined with an SVM classifier for the BI-RADS score which is a severity measure and part of the conclusion. The combined model had a ROUGE-L F1 score of 51.5% and an accuracy of 83.3% on the prediction of the BI-RADS score. The qualitative analysis showed that the model can generate comprehensible and relevant conclusions, while there is potential for improvement in the area of factual correctness.

Acknowledgment

We thank Rob Bourez for taking part in the qualitative evaluation as a radiologist at the ZGT hospital.

References

[1] E. A. Sickles, C. J. D'Orsi, and L. W. Basset, "Acr bi-rads® mammography," *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*, 2013.

[2] 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," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5 2019.

[3] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 1693–1701. [Online]. Available: http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf

[4] R. Nallapati, B. Zhou, C. dos Santos, Ç. GuÌ‡lçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. [Online]. Available: https://www.aclweb.org/anthology/K16-1028

[5] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, "Ontology-aware clinical abstractive summarization," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1013–1016. [Online]. Available: https://doi.org/10.1145/3331184.3331319

[6] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," 2017. [Online]. Available: https://arxiv.org/pdf/1704.04368

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[8] D. A. Sippo, G. I. Warden, K. P. Andriole, R. Lacson, I. Ikuta, R. L. Birdwell, and R. Khorasani, "Automated extraction of bi-rads final assessment categories from radiology reports with natural language processing," *Journal of digital imaging*, vol. 26, no. 5, pp. 989–994, 2013.

[9] S. M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, and R. S. Jacobson, "Automated annotation and classification of bi-rads assessment from radiology reports," *J. of Biomedical Informatics*, vol. 69, no. C, p. 177–187, May 2017. [Online]. Available: https://doi.org/10.1016/j.jbi.2017.04.011

[10] I. Banerjee, S. Bozkurt, E. Alkim, H. Sagreiya, A. W. Kurian, and D. L. Rubin, "Automatic inference of bi-rads final assessment categories from narrative mammography report findings," *Journal of biomedical informatics*, vol. 92, p. 103137, 2019.

[11] I. Mani, *Automatic Summarization*. John Benjamins Publishing Company, Jun. 2001. [Online]. Available: https://doi.org/10.1075/nlp.3

[12] D. Das and A. Martins, "A survey on automatic text summarization," 12 2007.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc, 2014, pp. 3104–3112.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. [Online]. Available: https://arxiv.org/pdf/1409.0473

[15] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *CoRR*, vol. abs/1509.00685, 2015. [Online]. Available: http://arxiv.org/abs/1509.00685

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[17] P. Gigioli, N. Sagar, A. Rao, and J. Voyles, "Domain-aware abstractive text summarization for medical documents," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2338–2343.

[18] B. Hu, A. Bajracharya, and H. Yu, "Generating medical assessments using a neural network model: Algorithm development and validation," 01 2020.

[19] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.

[20] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[21] C. Sammut and G. I. Webb, Eds., *TF–IDF*. Boston, MA: Springer US, 2010, pp. 986–987.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[24] D. Antognini, "py-rouge," 2018. [Online]. Available: https://pypi.org/project/py-rouge/

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019.

[26] S. Pathak, "Automatic structuring of breast cancer radiology reports for quality assurance," August 2018. [Online]. Available: http://essay.utwente.nl/76327/

[27] A. Johnson, T. Pollard, and R. Mark, "The mimic iii clinical database," 2015.

APPENDIX A.
RESULTS OF HYPERPARAMETER TUNING

Table IV
ENCODER-DECODER WITH ATTENTION (EDA) MODEL PERFORMANCE FOR DIFFERENT BATCH SIZES AND LATENT DIMENSIONS ON THE VALIDATION SET

Batch size	Latent dim.	BI-RADS accuracy	ROUGE-L			ROUGE-1			ROUGE-2		
			F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall
64	120	0.797	**0.519**	0.712	0.439	0.542	0.742	0.459	0.395	0.585	0.328
64	100	0.803	**0.518**	0.694	0.446	0.543	0.725	0.467	0.394	0.567	0.332
128	120	0.801	**0.517**	0.697	0.440	0.540	0.728	0.461	0.392	0.568	0.328
64	60	0.793	0.514	0.702	0.436	0.537	0.731	0.456	0.389	0.569	0.324
128	80	0.796	0.513	0.695	0.439	0.537	0.725	0.461	0.388	0.566	0.326
128	100	0.793	0.513	0.703	0.434	0.537	0.735	0.455	0.389	0.574	0.323
256	120	0.788	0.509	0.686	0.435	0.532	0.716	0.455	0.383	0.555	0.321
128	60	0.773	0.508	0.683	0.435	0.531	0.714	0.456	0.381	0.548	0.320
64	80	0.759	0.506	0.705	0.426	0.525	0.731	0.443	0.379	0.566	0.313
256	100	0.767	0.504	0.695	0.428	0.528	0.726	0.449	0.377	0.558	0.314
512	100	0.769	0.502	0.692	0.427	0.525	0.721	0.446	0.377	0.556	0.314
512	60	0.772	0.501	0.675	0.429	0.523	0.704	0.449	0.373	0.538	0.314
256	60	0.774	0.501	0.686	0.424	0.522	0.713	0.443	0.374	0.549	0.311
512	120	0.768	0.500	0.662	0.435	0.523	0.691	0.456	0.371	0.528	0.316
256	80	0.782	0.496	0.693	0.419	0.519	0.722	0.438	0.369	0.555	0.305
512	80	0.750	0.494	0.692	0.415	0.515	0.720	0.433	0.365	0.550	0.301