university of groningen

University Medical Center Groningen

University of Groningen

The natural language processing of radiology requests and reports of chest imaging

Olthof, Allard W.; van Ooijen, Peter M.A.; Cornelissen, Ludo J.

Published in:
Health Informatics Journal

DOI:
10.1177/14604582221131198

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Olthof, A. W., van Ooijen, P. M. A., & Cornelissen, L. J. (2022). The natural language processing of radiology requests and reports of chest imaging: Comparing five transformer models' multilabel classification and a proof-of-concept study. *Health Informatics Journal*, *28*(4). https://doi.org/10.1177/14604582221131198

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 20-11-2022

Check for updates

Original Research Article

Health Informatics Journal

Health Informatics Journal
1–26
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14604582221131198
journals.sagepub.com/home/jhi
SAGE

The natural language processing of radiology requests and reports of chest imaging: Comparing five transformer models' multilabel classification and a proof-of-concept study

Allard W Olthof
Department of Radiology, Hospital Group Twente, Almelo, Netherlands

Peter MA van Ooijen
Department of Radiation Oncology, University Medical Center Groningen, Netherlands
Data Science Center in Health (DASH), University Medical Center Groningen, Netherlands

Ludo J Cornelissen
Department of Radiation Oncology, University Medical Center Groningen, Netherlands
COSMONiO Imaging B.V., Groningen, Netherlands

Abstract

Background: Radiology requests and reports contain valuable information about diagnostic findings and indications, and transformer-based language models are promising for more accurate text classification.

Methods: In a retrospective study, 2256 radiologist-annotated radiology requests (8 classes) and reports (10 classes) were divided into training and testing datasets (90% and 10%, respectively) and used to train 32 models. Performance metrics were compared by model type (LSTM, Bertje, RobBERT, BERT-clinical, BERT-multilingual, BERT-base), text length, data prevalence, and training

Corresponding author:
Allard W Olthof, Department of Radiology, Hospital Group Twente, Zilvermeeuw 1, Almelo 7609 PP, Netherlands.
Email: a.olthof@zgt.nl

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

strategy. The best models were used to predict the remaining 40,873 cases' categories of the datasets of requests and reports.

Results: The RobBERT model performed the best after 4000 training iterations, resulting in AUC values ranging from 0.808 [95% CI (0.757–0.859)] to 0.976 [95% CI (0.956–0.996)] for the requests and 0.746 [95% CI (0.689–0.802)] to 1.0 [95% CI (1.0–1.0)] for the reports. The AUC for the classification of normal reports was 0.95 [95% CI (0.922–0.979)]. The predicted data demonstrated variability of both diagnostic yield for various request classes and request patterns related to COVID-19 hospital admission data.

Conclusion: Transformer-based natural language processing is feasible for the multilabel classification of chest imaging request and report items. Diagnostic yield varies with the information in the requests.

Keywords
natural language processing, machine learning, data mining, radiology, chest imaging

Introduction

Radiology reports are the primary communication method between radiologists and referring physicians and contain valuable information that impacts patient care.[1,2] Most radiology reports are for single use, and physicians use the information from them to treat their patients. Data science can prove valuable by providing information from aggregated data; this is certainly true for radiology reports.[3]

Radiology dashboards are suitable for monitoring and predicting radiology volumes and resource utilization;[4,5] however, these systems do not provide in-depth information about referrals or diagnostic findings. Therefore, these systems are less suitable for providing insight into aggregate data about, for example, imaging appropriateness.[6] Reasons for referrals (or the input for the radiology process) can be retrieved from the content of the requests.[7] The percentage of positive findings, the diagnostic yield, can be calculated from the information in the radiology report. The diagnostic yield provides insight into disease prevalence, which not only informs referring clinicians but also can impact patient management.[8] Insight into appropriateness (input) and diagnostic yield (output) is not routinely assessed, even though this type of information is valuable—it contributes to effective resource utilization[9–12] and allows for the identification of factors related to overtesting.[13]

To leverage the great potential of aggregated data, the retrieval of information from radiology requests and reports must be automated. This is particularly the case when documents need to be classified for multiple co-occurring items. Instead of obtaining this information by manually categorizing radiology requests and reports, data mining using natural language processing (NLP) offers a promising alternative.[14,15] In the field of chest imaging, traditional NLP has been successfully applied and can identify pulmonary nodules,[16] pneumonia,[17] and pulmonary embolisms[18] from radiology reports. Deep learning-based NLP methods perform equally well or better than traditional NLP models and can also be used to classify chest radiology reports.[19] High-performance transformer-based NLP algorithms, such as Bidirectional Encoding Representations for Transformers (BERT), have been applied to medical texts and are available open-source in several languages.[20–22] The strengths of transformer-

based NLP models are that they are pretrained on large datasets and account for a word's context.[20,23] These pretrained models can be applied to various NLP tasks after fine-tuning; that is, additional training with data from a specific task. Transformer-based models have been used for different NLP-tasks in radiology such as report section segmentation,[24] assessing spational information in radiology reports,[25] dectection of actionable findings,[26] image annotation,[27] and radiology report generation.[28]

In our previous work, we proved BERT's superior performance during a single-label classification task compared to other deep learning NLP methods, such as the convolutional neural network (CNN) and the long short-term memory (LSTM) neural network.[29] To our knowledge, the application of transformer-based models for the multilabel classification of a combination of chest imaging requests and reports (which can contribute to increased insights into the role and performance of chest imaging by assessing the diagnostic yield of different categories) has not yet been reported by other authors. Both the development and evaluation of a multilabel classification method are prerequisites for further research about the application of NLP to the evaluation of clinical care and the leveraging of data for predictive purposes.

In the current study, we hypothesise that (1) transformer-based NLP models that have been pretrained with Dutch language data will perform better compared to multilingual or English models when fine-tuned for radiology requests and reports written in Dutch; (2) multilabel classification with transformer-based NLP will perform equally well among different request and report categories and will be comparable to single-label classification; and (3) diagnostic yield will vary depending on the information in the request. Our research objectives are as follows:

1. Develop a deep learning-based NLP pipeline for the multilabel classification of radiology requests and reports.
2. Apply the pipeline to train and test five transformer-based models from a chest imaging dataset sample, and then compare performance metrics pertaining to model type, training method, and text characteristics.
3. Use the best-trained model to predict the classifications of requests and reports to demonstrate the feasibility of applying NLP to an extensive dataset in a proof-of-concept study of chest imaging, and analyse the relationships among the referral reason, diagnostic yield, and variability of requests over time.

Methods

An NLP pipeline was developed in a Python Jupyter notebook with sections for data import, training, testing, prediction, data analysis, and visualization (Table 1).

In a retrospective study, we used annotated datasets of radiology requests and reports to train and test five transformer-based NLP models for multilabel classification: BERTje,[30] RobBERT,[31] BERT-multilingual,[32] BERT-clinical,[33] and BERT-base.[23] An LSTM model was trained as a baseline for comparison.

The best-performing models were used to predict the classification of an unannotated dataset of requests and reports. According to Dutch law regarding medical research on humans, no informed consent was needed because of the nature of the retrospective chart review. The project was approved by the local research committee and the hospital's board of directors.

Table 1. The NLP pipeline, organised into sections of Python code.

Data import	
Requests	Request texts and labels extracted and copied to a separate data frame; randomization and construction of training (90%) and testing (10%) datasets
Reports	Report texts and labels extracted and copied to a separate data frame; randomization and construction of training (90%) and testing (10%) datasets
Dashboard	All data with additional variables indicate the presence or absence of annotations for reports and requests; different data frames with unannotated requests and reports are for prediction
Requests pipeline	
Training	Model training on the request training dataset
Testing	Model testing on the request testing dataset
Reports pipeline	
Training	Model training on the report training dataset
Testing	Model testing on the report testing dataset
Analysis pipeline	
Predictions	Creation of a data frame with predicted labels; copying predicted labels for reports and requests to the analysis data frame
Visualization	Charting of annotated dataset distribution, model performance metric, use cases for complete annotated and predicted datasets

Data

All requests for and reports of chest radiographs (CR) and computed tomography (CT) from January 2020 to September 2021 (n = 43,129) were retrieved from the PACS of a general hospital in the northern part of the Netherlands (Treant Healthcare Group).

The dataset was pseudonymized and stored in both the hospital IT system and secure cloud storage. Figure 1 shows the data processing flowchart.

Ground Truth

In multilabel classification, each text item has multiple binary labels. For example, a chest imaging request that says, "*Patient with fever and coughing. Infiltrate? Pleural fluid? Signs of COVID-19?*" has positive labels for "Infiltrate", "Pleural fluid", and "COVID-19", but negative labels for other items such as "Tumor". In this manner, a model can be trained to classify texts using multiple labels simultaneously. For model training and testing, the requests (n = 2256) and reports (n = 2256) from the first 2 weeks of March 2020 and the first 2 weeks of April 2020 were annotated by a board-certified radiologist with 13 years of chest imaging experience. The two time periods were considered representative of cases without and with COVID-19, respectively, because the local rise of COVID-19 cases occurred towards the end of March 2020. The annotation process, performed in Microsoft Excel, consisted of assigning eight nonexclusive categories to the requests and 10 nonexclusive categories to the reports. The categories were based on the most frequent items found in the requests and the most frequent findings in the reports. Items had to be explicitly mentioned to be labelled as positive. For example, the COVID-19 label was given to requests that explicitly mentioned COVID-19 or coronavirus. For the reports, the "Normal" category was used only for those devoid of any abnormal findings.

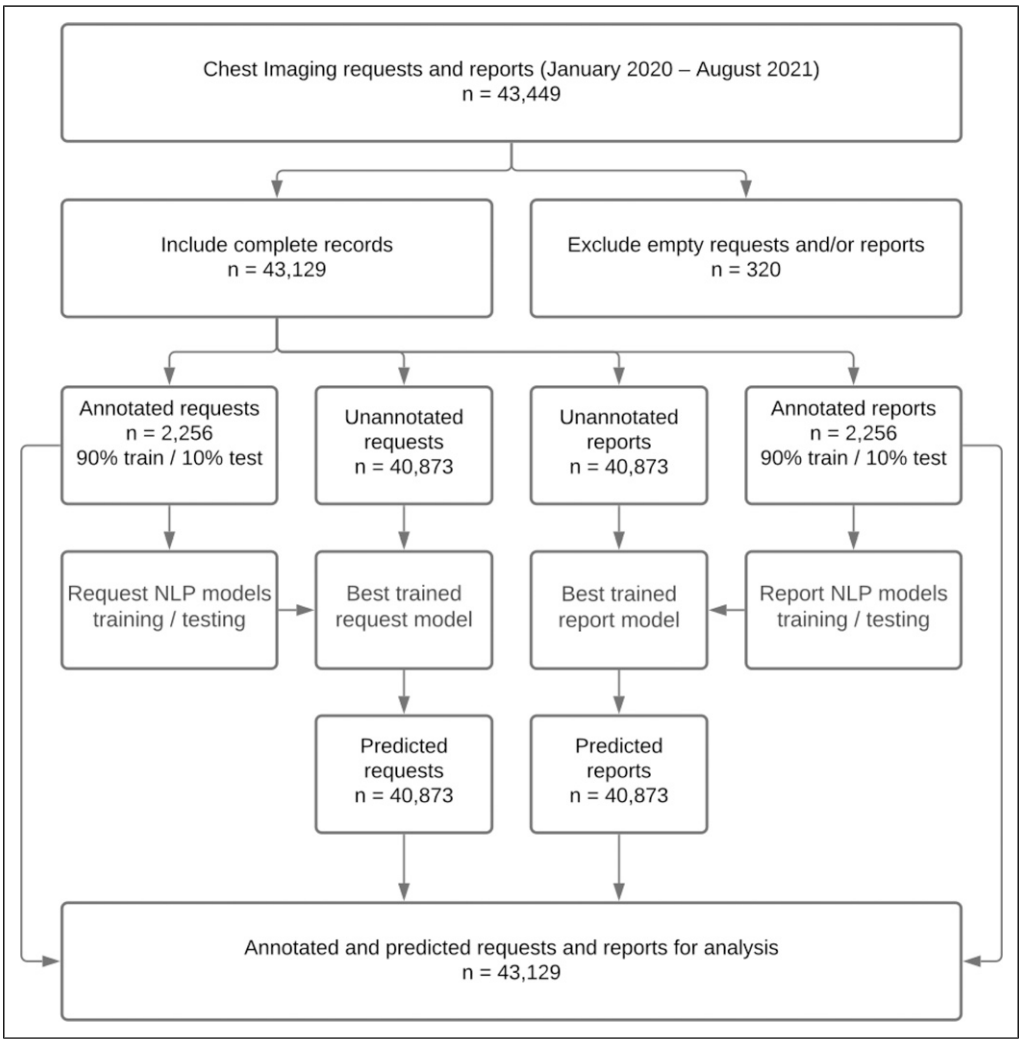

Figure 1. Data processing flowchart.

After 5 months, the same radiologist reviewed all annotations to ensure consistency before the final model training was performed. This quality check resulted in the correction of 86 errors with the request annotations (4%) and 59 errors with the report annotations (3%).

Data Partitions

The annotated data was split into two groups, training data (90%) and testing data (10%), by using iterative_train_test_split from the skmultilearn library and used for model development.[34,35] The remaining unannotated data (n = 40,873) from the original dataset was sent through the best-performing request and report models to receive classification predictions. The combined data (i.e. the combination of the annotated and predicted data) was used for the proof-of-concept study.

Models

The NLP library Simple Transformers (simpletransformers.ai) was used to access the Hugging Face library,[36] in which different pretrained transformer-based NLP architectures are available, including BERT-based models.[23] The models used for multilabel classification are specified in Table 2.

Acting as a baseline, an LSTM model was trained according to the methodology described in our previous work[29] except for the eight request and 10 report classes used for multilabel (instead of single-label) classification of the last layer.

Training

The training parameters used are defined in Table 3. The request and report models were trained for 16 epochs. The request models had a maximum sequence length of 128, and the report models had a maximum length of 512. Class weights that were inversely proportional to class frequencies were applied to reduce the impact of class imbalances. During the training after each epoch, the models were evaluated for accuracy. The overall best-performing models were identified, overwrote the poorly performing models, and were saved. (The LSTM models were trained for 3, 8, 16, 32, 64, 128, and 256 epochs; the best-performing request and report models were included in the evaluation).

Evaluation

For each of the five model types, three trained models—the models with the best accuracy according to the evaluation during training, the trained model after 2000 iterations, and the trained model after 4000 iterations—were stored and used for evaluation. This approach was implemented to identify the impact of training duration on performance. For the 15 trained transformer models and the baseline LSTM models using the testing sets, model performance was evaluated by calculating sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the curve (AUC), and F1-score. This resulted in 128 and 160 sets of performance metrics for request and report data, respectively. Confidence intervals were calculated. Using box plots, the following factors were compared to evaluate their impact on performance:

1. Model type (LSTM vs. Bertje vs. RobBERT vs. BERT-clinical vs. BERT-multilingual vs. BERT-base),
2. Data type (short requests vs. long reports),
3. Data prevalence (separate analysis of the request and report items), and
4. Training strategy (overall best vs. 2000 iterations vs. 4000 iterations).

Statistical significance was assessed using Python and the SciPy Library. An uncorrected p-value of <0.05 was considered statistically significant.

Proof-of-concept Study

To identify patterns on a large scale, a proof-of-concept study was performed with automated annotations of the unlabelled data. Thus, all the predicted request and report data ($n = 40{,}873$) was

Table 2. Transformer model characteristics. Transformer models are language models that have been pretrained on text corpora in one or more languages. Listed here are the characteristics of the study's five models. Fine-tuning was performed on Dutch radiology requests and reports.

	BERTje	RobBERT	BERT_multilingual	Clinical_BERT	BERT_base
Modeltype	bert	Roberta	bert	bert	bert
Modelname	GroNLP/bert-base-Dutch-cased	Pdelobelle/robbert-v2-Dutch-base	Bert-base- multilingual-cased	emilyalsentzer/Bio_ClinicalBERT bert-base-cased	
Language	Dutch	Dutch	Multiple, including Dutch	English	English
Self attention layers	12	12	12	12	12
Heads per layer	12	12	12	12	12
Trainable parameters	110 million	117 million	110 million	110 million	110 milion
Pretraining data	Books, TwNC (Dutch news corpus), SoNaR-500 (multigenre reference corpus), Web news, Wikipedia	Dutch section of OSCAR (Open super large crawled aggregatred corpus)	Wikipedia (top 102 languages with the largest Wikipedia)	All notes from MIMIC III, a database containing electronic health records from ICU patients at the Beth Israel hospital in Boston, MA	BooksCorpus, English Wikipedia
Data size (words)	2.4 billion	6.6 billion	non specified	880 million (after being pretrained on BioBERT 21.3 billion words)	3.3 billion
Website	https://huggingface.co/GroNLP/bert-base-Dutch-cased	https://huggingface.co/pdelobelle/robbert-v2-Dutch-cased	https://huggingface.co/bert-base multilingual cased	https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT	https://huggingface.co/bert-base-cased

Table 3. Model training parameters of the MultiLabelClassificationModel from the simpletransformers library. Parameters can be changed to optimise the training process for a specific situation.

Parameter	Value
num_train_epochs	16
evaluate_during_training	True
evaluate_during_training_verbose	True
overwrite_output_dir	True
save_model_every_epoch	False
use_early_stopping	True
early_stopping_delta	0.0005
early_stopping_patience	3
num_labels (requests)	8
num_labels (reports)	10
use_cuda	False
learning_rate	4.00×10^{-5}
Optimizer	AdamW
train_batch_size	8
save_steps	2000

entered into the best-performing trained request and report models, respectively, to obtain predicted labels.

The variable number of items for each request and report equalled the sum of the positive labels per request or report. For all request categories, the likelihood of positive findings in the corresponding reports was calculated using contingency tables of each combination of requests and report categories. A heatmap was created that showed the diagnostic yield of the reported item for all single-item requests.

The variation of request items per week was visualised and combined with hospital admission data (retrieved from a publicly available dataset[37]) originating from the same region as the radiology data.

Results

Data

Figure 2 displays the number of requests and reports included in the study, as well as each file's word count, which varied among modality type (CR or CT) and document type (request or report). For example, the requests' word count was the lowest. CR requests were the smallest texts, and the largest texts were CT reports.

Table 4 presents the training and testing datasets' categories and their prevalence. Both datasets had the same distribution. The class imbalance across the labels is apparent for both the requests and the reports. For the requests, the categories "Infiltrate" (31.8%), "Tumor" (20.5%), and "Other" (47.4%) were seen most frequently. For the corresponding reports, the percentages of positive findings were smaller than those of the requests. "Other" (69.1%), "Normal" (21.7%), and "Infiltrate" (15.8%) were the most frequent positive findings.

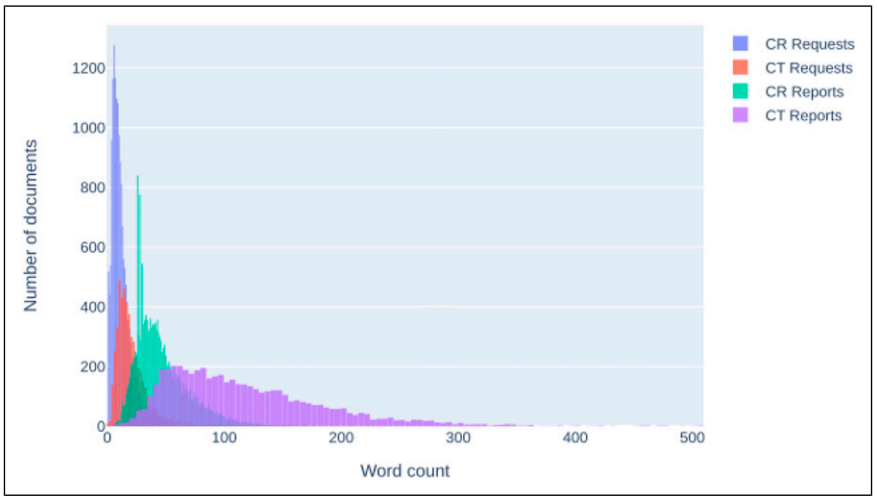

Figure 2. Number of and word count for radiography (CR) and computed tomography (CT) requests and reports.

Table 4. Prevalence of report (a) and request (b) items in the training and testing datasets. For each case, one or more items could be absent (0) or present (1). The frequency and percentage of positive labels in the training set and test set were calculated. Per definition, the sum of items exceeds the total number of cases because, in multi-label annotation, multiple items can co-occur.

(a)	Train	Train %	Test	Test %	*p*-value
Report_Infiltrate	321	15.9	36	15.5	0.999782
Report_Decompensation	48	2.4	5	2.2	
Report_Tumor	180	8.9	20	8.6	
Report_Pleural_fluid	229	11.3	26	11.2	
Report_Pulmonary_embolism	14	0.7	2	0.9	
Report_Pneumothorax	31	1.5	4	1.7	
Report_Other	1404	69.4	156	67.2	
Report_Groundglass	80	4	9	3.9	
Report_COVID19	41	2	6	2.6	
Report_Normal	440	21.7	49	21.1	
(b)	Train	Train %	Test	Test %	*p*-value
Request_Infiltrate	646	31.8	72	32.3	0.999999
Request_Decompensation	163	8	18	8.1	
Request_Tumor	417	20.5	46	20.6	
Request_Pleural_fluid	159	7.8	18	8.1	
Request_COVID19	190	9.3	21	9.4	
Request_Pulmonary_embolism	76	3.7	8	3.6	
Request_Pneumothorax	154	7.6	17	7.6	
Request_Other	963	47.4	107	48	

Table 5. Comparisons of AUC values in different combinations of models (a), training durations (b), item prevalence (c), and datasets (d). If two categories within a group were compared, *p*-values were calculated; categories that were found to be statistically significantly better are indicated by bold–italic font.

Dataset	Category 1	mean1	Category 2	mean2	tstat	Pvalue
	A. Models					
Requests	BERTje	0.85	***RobBERT***	0.89	−2.6188013	***0.0101***
Requests	BERTje	0.85	BERT_multilingual	0.87	−1.174089	0.243
Requests	BERTje	0.85	Clinical_BERT	0.85	−0.0394513	0.9686
Requests	BERTje	0.85	BERT_base	0.84	0.5062153	0.6138
Requests	RobBERT	0.89	BERT_multilingual	0.87	1.37262385	0.1728
Requests	***RobBERT***	0.89	Clinical_BERT	0.85	2.15532423	***0.0334***
Requests	***RobBERT***	0.89	BERT_base	0.84	2.63919662	***0.0096***
Requests	BERT_multilingual	0.87	Clinical_BERT	0.85	0.94373342	0.3475
Requests	BERT_multilingual	0.87	BERT_base	0.84	1.45008372	0.15
Requests	Clinical_BERT	0.85	BERT_base	0.84	0.47495099	0.6358
Reports	BERTje	0.85	***RobBERT***	0.89	−2.6188013	***0.0101***
Reports	BERTje	0.85	BERT_multilingual	0.87	−1.174089	0.243
Reports	BERTje	0.85	Clinical_BERT	0.85	−0.0394513	0.9686
Reports	BERTje	0.85	BERT_base	0.84	0.5062153	0.6138
Reports	RobBERT	0.89	BERT_multilingual	0.87	1.37262385	0.1728
Reports	***RobBERT***	0.89	Clinical_BERT	0.85	2.15532423	***0.0334***
Reports	***RobBERT***	0.89	BERT_base	0.84	2.63919662	***0.0096***
Reports	BERT_multilingual	0.87	Clinical_BERT	0.85	0.94373342	0.3475
Reports	BERT_multilingual	0.87	BERT_base	0.84	1.45008372	0.15
Reports	Clinical_BERT	0.85	BERT_base	0.84	0.47495099	0.6358
	b. Training duration					
Requests	2/3 epochs, 510/765 iterations	0.86	7.8 epochs, 2000 iterations	0.86	−0.131003	0.8959
Requests	2/3 epochs, 510/765 iterations	0.86	15.7 epochs, 4000 iterations	0.85	0.39539041	0.693
Requests	7.8 epochs, 2000 iterations	0.86	15.7 epochs, 4000 iterations	0.85	0.53186221	0.5955
Reports	2/3 epochs, 510/765 iterations	0.86	7.8 epochs, 2000 iterations	0.86	−0.131003	0.8959
Reports	2/3 epochs, 510/765 iterations	0.86	15.7 epochs, 4000 iterations	0.85	0.39539041	0.693
Reports	7.8 epochs, 2000 iterations	0.86	15.7 epochs, 4000 iterations	0.85	0.53186221	0.5955
	c. Item prevalence					
Requests	***high***	0.89	low	0.84	4.00709611	***0.0001***
Reports	***high***	0.89	low	0.84	4.00709611	***0.0001***
	d. Dataset					
Requests and reports	***Requests***	0.9	Reports	0.81	7.57224074	***<0.0001***

Model Performance

Training and testing. The models' performance metrics from the training dataset are summarised in Figures 3 and 4. Raw data of the model performance metrics with 95% confidence intervals is provided in the Supplemental appendix (Table A1).

The metrics for the models trained on the request dataset were better than the report dataset models. In addition, the latter showed considerably more variation in performance. This variation occurred between the models and the training strategies. Especially true for the reports, the transformer models that were pretrained with Dutch data performed better, needing only 2–3 training epochs and improving only modestly after additional training. The specificity of the multilingual and English pretrained models improved substantially with additional training, but the AUC did not change significantly when training strategies were compared because the sensitivities decreased after prolonged training.

For both requests and reports, each model demonstrated a high specificity (> 0.90) and negative predictive value (> 0.95). Also, the sensitivities and positive predictive values were more variable and greater for the request and report items of higher prevalence in the datasets.

Combinations of AUC values and F1-scores are compared for statistical significance in Table 5 and Table 6.

The transformer models outperformed the baseline LSTM models, but the RobBERT model surpassed all the others. The RobBERT model that trained for 4000 iterations was chosen as the best model overall. For the requests, the AUC values varied from 0.808 [95% CI (0.757–0.859)] for the

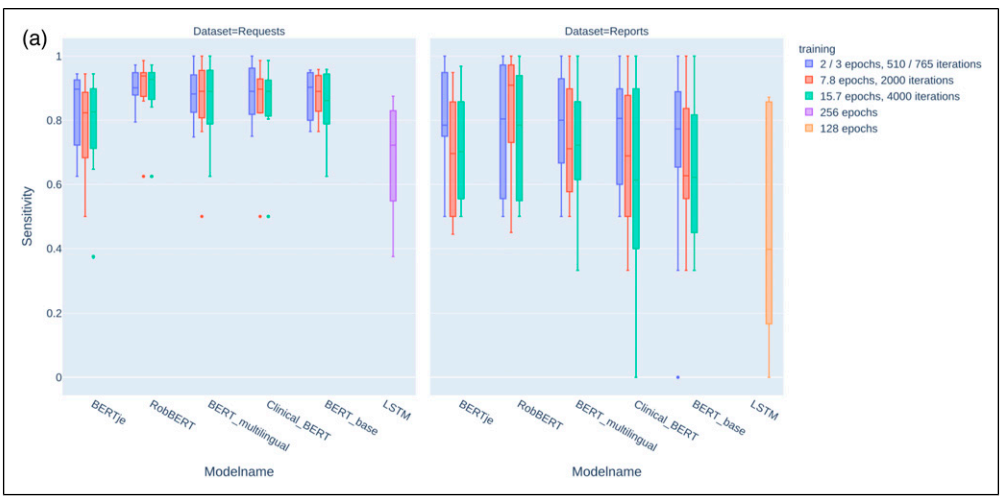

Figure 3. Summarised performance metrics of five transformer models and one LSTM model. For each model type, three models with different training durations were evaluated. Colours indicate training duration: two or three epochs (according to evaluation during training), 7.8 and 15.7 epochs for the transformer-based models, and 256 and 128 epochs for the LSTM models (best performance empirically). The sensitivity (a), specificity (b), positive predictive value (c), negative predictive value (d), AUC (e), and F1-score (f) were calculated for all labels and displayed as box plots per model type.

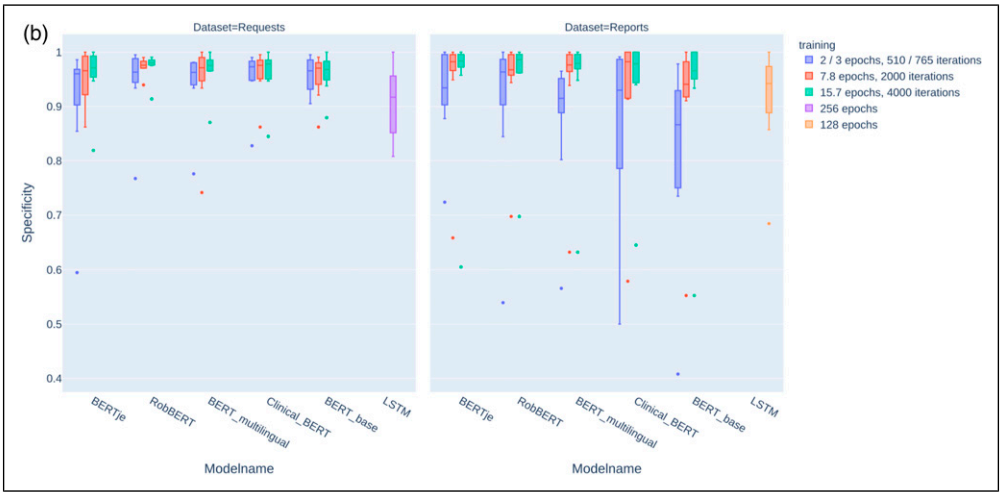

Figure 3. continued.

item Request_Pulmonary_embolism to 0.976 [95% CI (0.956–0.996)] for the item Request_infiltrate. For the reports, the AUC values varied from 0.746 [95% CI (0.689–0.802)] for the item Report_COVID-19 to 1.0 [95% CI (1.0–1.0)] for the (infrequent) item Report_Pneumothorax. The AUC for the classification of normal reports was 0.95 [95% CI (0.922–0.979)]. In the comparison of F1-scores the differences between models were less pronounced, but did show significantly better results for longer training duration. Both the comparisons of AUC and F1-score demonstrated larger differences between results of data prevalence and text size compared to differences between models.

Predictions and proof-of-concept study. Most requests had one item (26,047; 64%) or two items (10,448; 26%). Requests with 3 (3587; 8%), 4 (789; 2%) or more items were much less frequent. In Table 7, the likelihood of report items is depicted, given the presence of specific request items. For each request item, there was a variable yield of report findings, not only for the item in the request but also for other items.

Figure 5(a) and Figure 5(b) illustrates the variability of the diagnostic yield per request category. For example, infiltrates are found when they are specifically mentioned in requests, but they can also be found even if requests specify other categories.

Figure 6 shows the variability in request volume over time. Radiology data can be combined with data from other sources; in this case, the rise of COVID-19 requests corresponds to an increase in hospital admissions. The first wave of COVID-19 led to an overall reduction of imaging, including requests in the "Infiltrate" category.

Discussion

In this study, we developed a pipeline for deep learning NLP in the context of radiology and compared five transformer models and one LSTM model. Distinctive characteristics of our work are

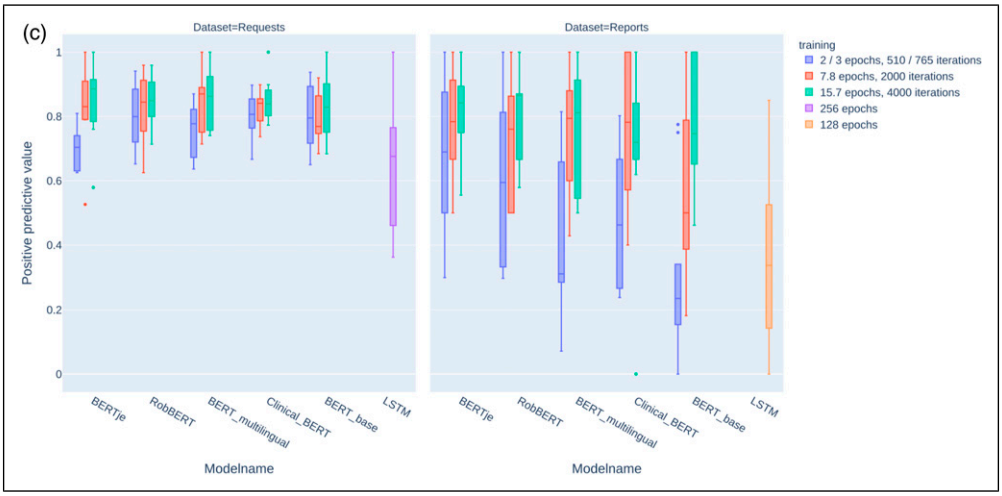

Figure 3. continued.

the number of different models that we compared, as well as the multilabel (instead of single-label) classification of both radiology reports and requests. We now discuss (1) the possible reasons for the obtained results, (2) this work's contribution and its comparison with published research, and (3) the challenges and limitations of this study.

Explanation of Results

The major difference between the LSTM models and the transformer models was that the former was trained only on the training data and the latter was pretrained on large corpora and took the context of words into account. This meant that, in our study, the transformer models needed less training with the training data to reach high performance levels. The Bertje and RobBERT models were pretrained with Dutch text, and not only was their better performance with short training duration expected, but it also confirmed one of our hypotheses. The multilingual and English models' performance improved substantially after longer training, indicating the adaptability of transformer models. However, for pretraining, both language and type of text are relevant. For example, BERT-clinical was trained with English medical texts and, consequently, performed better than BERT-base. The results did not empirically explain the superior performance of the RobBERT model, but the model characteristics provide clues that can explain the differences. First, the RobBERT model was pre-trained on a corpus in the Dutch language, and second, the size of this corpus surpassed that of the other pre-trained Dutch model.

Prevalence impacts model performance, especially regarding sensitivity and positive predictive value. Accordingly, classification metrics change because of the differences in the prevalence of request and report labels. The datasets' class imbalances were greater for the reports than for the requests; they were probably not fully compensated for by the application of class weighting. The word counts (and variations in word count) of the reports were greater than those of the requests,

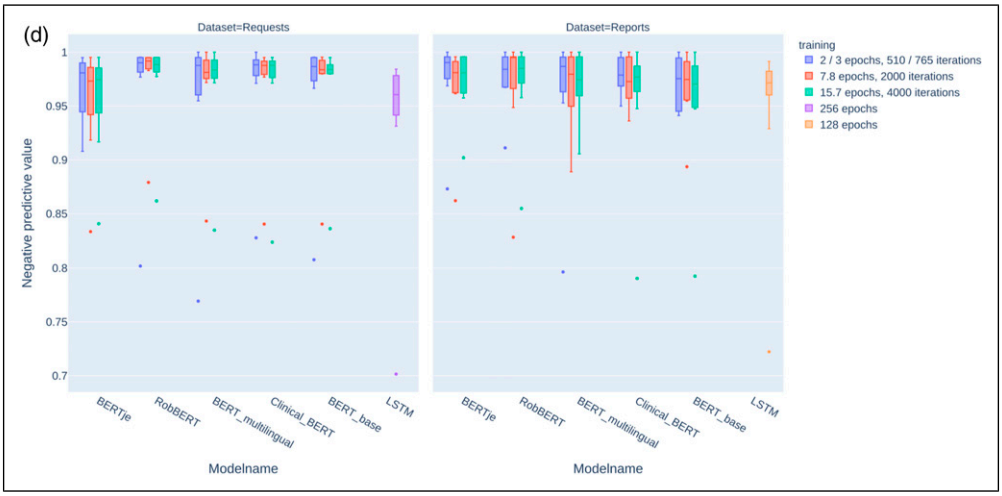

Figure 3. continued.

which explains the higher performance of the request models compared to the report models. This variability of performance, dependent on text characteristics and prevalence, confirmed our hypothesis about the comparable performance of multilabel classification compared to the single-label classification of our previous work.[29] The study results confirm previous results that transformer algorithms have higher performance in classification task on shorter texts and text with less class imbalance.[29]

The proof-of-concept study illustrated and confirmed our hypothesis regarding diagnostic yield variations among different request categories. The degree of variation was difficult to estimate beforehand: just as one disease can cause several abnormalities, this was reflected in the radiology report findings. Therefore, multilabel classification provides a better reflection of the data than single-label classification.

Contributions and Comparison with the Literature

Natural language processing is increasingly applied to radiology reports,[15] but comparable studies of transformer-based models applied to chest imaging are scarce. Our work adds to the existing evidence that transformer models can reliably classify radiology reports and demonstrates the feasibility of combining both request and report classification models in the same pipeline. In addition, the study demonstrated the impact of dataset characteristics, such as item prevalence and text length. The ability of transformer models to classify both radiology requests and reports is important because the information provided by NLP models enables large-scale data retrieval for a myriad of other downstream tasks, such as analysing imaging results over extensive periods, as demonstrated in the proof-of-concept study. This work, therefore, contributes to the notion that NLP just as other applications of artificial intelligence can augment the ability of radiologists in patient care.[38]

Wood et al. applied BioBERT to neuroradiology reports and achieved superior performance compared to our study.[39] It is important to remember that BioBERT combines language and medical

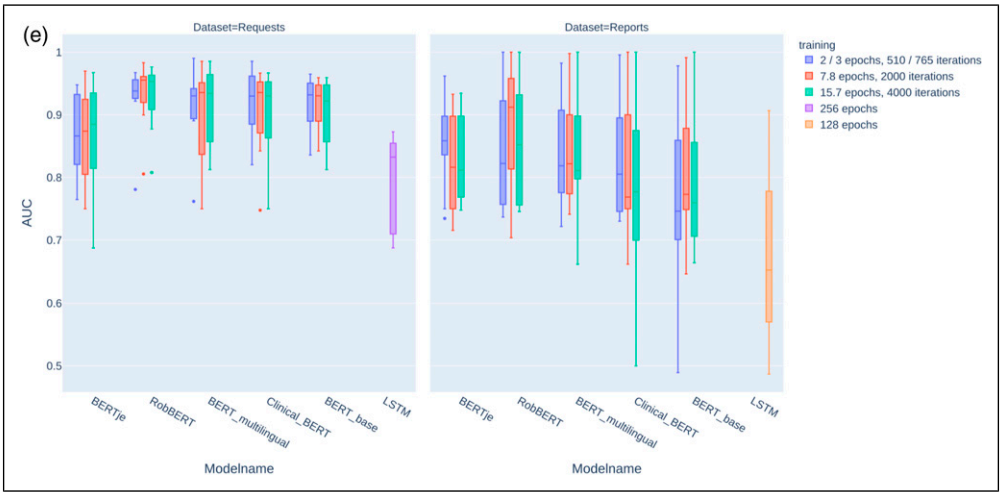

Figure 3. continued.

context in the pretraining dataset; however, both studies saw differences in performance dependent on class labels.

Bressem et al.[40] compared four BERT models' classifications of chest radiography reports. The best-performing model achieved the best pooled AUC of 0.98. The higher performance compared to our study can be explained by the inclusion of chest radiographs from intensive care patients, which would increase the prevalence of dataset findings, and the exclusion of reports that provided information about a single item without mentioning the absence or presence of other items. Our dataset is, therefore, less homogenous, but better reflects data in daily practice. Preselection can improve results, but this type of exclusion must be taken into account when applying the trained model to new data. The omission of preselection allowed us to use the trained models for predictive purposes on an unseen dataset.

Venturelli et al. assessed the appropriateness of referrals for imaging and other diagnostic procedures by analysing the requests' content using a commercial software package.[41] Their study's focus was on classification results (appropriate vs. not appropriate) and not the applied method's performance, which impeded comparison with our study. However, the study demonstrated the feasibility of using requests for assessing appropriateness and is an excellent example of a future application for transformer-based NLP.

A systematic review of studies regarding the diagnostic yield of head CT scans of patients with syncope declared that a small sample size was a limitation of many studies and advocated large prospective research.[42] Our pipeline can be applied to such research in various subspecialties because of its applicability to large datasets. Pons et al. applied NLP to the long-term evaluation of the diagnostic yield of head CT for patients with minor head injuries.[43] Similar to our study was the use of information about the indications of imaging and information about the diagnostic results. The authors also used a small part of the data to train a model to extract information from a larger dataset.

Annarumma et al.[44] applied NLP to annotate chest radiographs to train a deep-learning image classification model for critical, urgent, nonurgent, and normal categories. The F1-scores for the extraction of the presence or absence of radiologic findings within the free-text

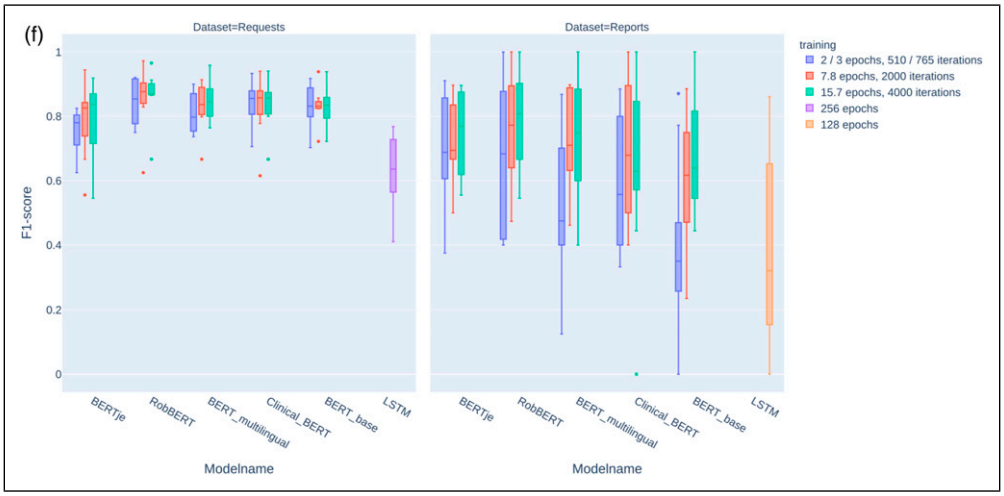

Figure 3. continued.

reports ranged from 0.81–0.99 and were in the same range as our study. Annarumma et al. used a rule-based approach compared to deep learning in our study. The advantage of a deep learning pipeline is that it can be used for other annotated datasets; in contrast, a rule-based method requires manual feature engineering. Niehues reported a computer vision model for chest radiographs that was trained with labels derived from transformer-based NLP and demonstrated promising results comparable to those from expert radiologists.[27]

The exceptional performance of the multilabel classification of radiology reports is not unique for transformer models. Short et al. reported a multilabel classification of mammography reports[45] and compared rule-based methods with a combination of convolutional and recurrent neural networks. Besides the methods applied, other differences between our studies included the relatively structured, homogeneous data used and the word-level (instead of document-level) classification performed.

Challenges and Limitations

No formal assessment of randomness of data or evaluation metrics was performed. Because all models were trained with the same data set, this is not supposed to have an impact on the results. Another limitation is that during training for all models all parameters were kept constant. Further performance optimization could be possible by hyperparameter tuning.

As already mentioned, model performance was impacted by class imbalances. This was relevant for both the training and testing datasets. Sensitivity values would have been more robust with a larger testing dataset size. Future research should incorporate additional methods to overcome low sensitivity due to class-imbalance.[46] Alternative performance measures can be considered in case of class-imbalance.[47] Furthermore, manual annotation was performed by a

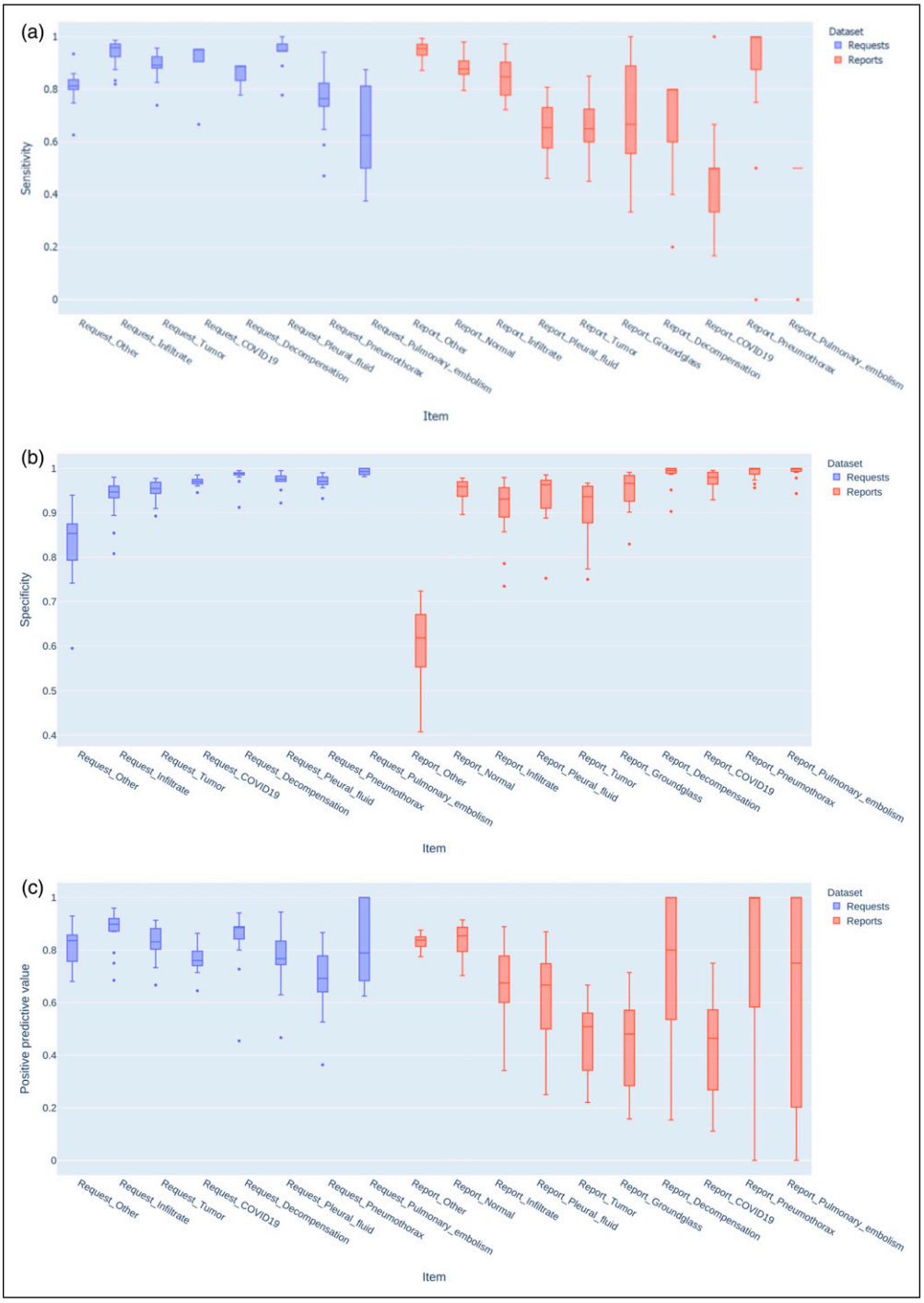

Figure 4. Summarised performance metrics for all request and report items, as well as all models and training schedules. Request and report items are sorted in descending order, from left to right, according to prevalence.

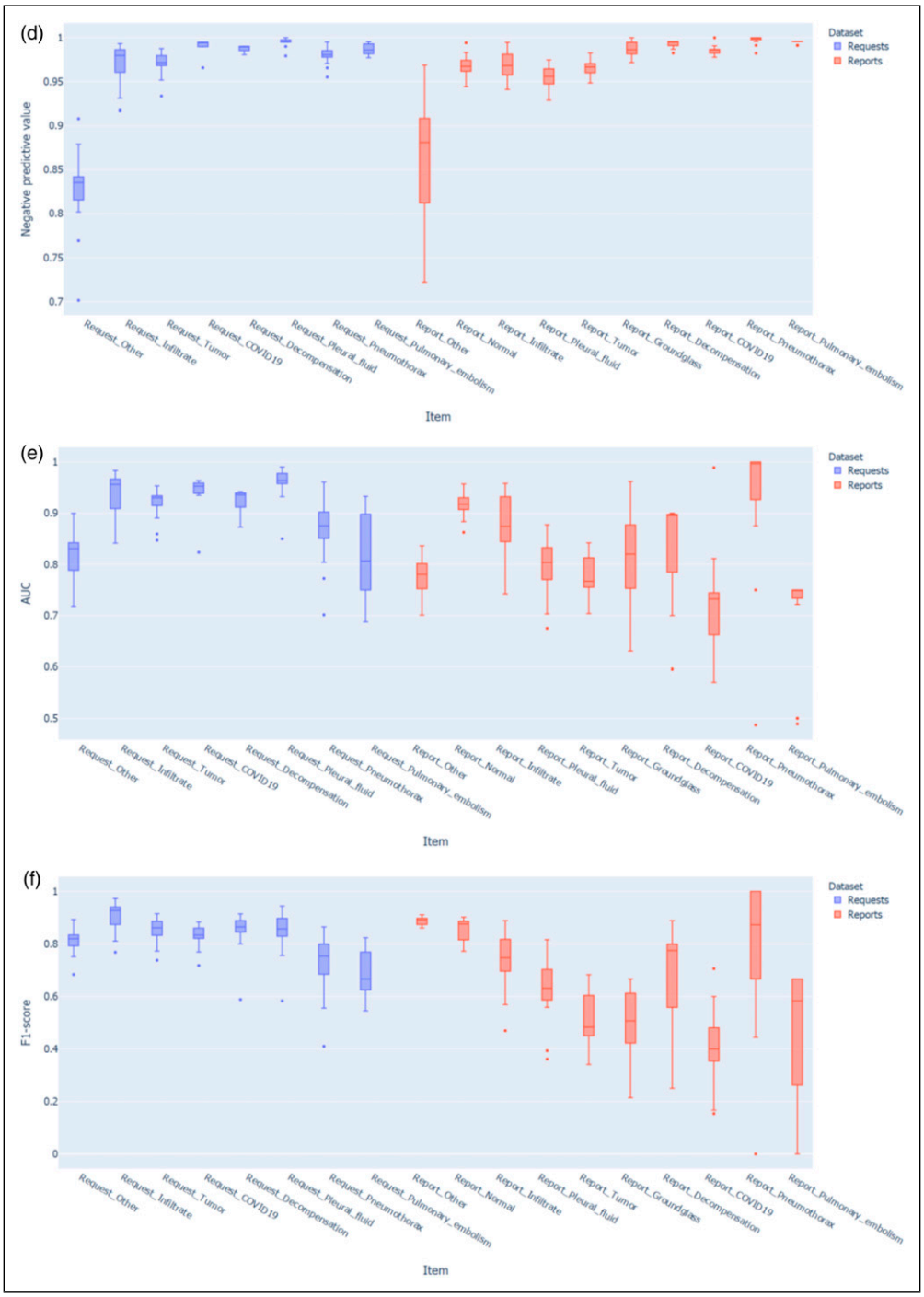

Figure 4. continued.

Table 6. Comparisons of F1 scores in different combinations of models (a), training durations (b), item prevalence (c), and datasets (d). If two categories within a group were compared, *p*-values were calculated; categories that were found to be statistically significantly better are indicated by bold–italic font.

Dataset	categ1	mean1	categ2	mean2	tstat	Pvalue
	a. Models					
Requests	BERTje	0.78	***RobBERT***	0.85	−2.743105605	***0.0086***
Requests	BERTje	0.78	BERT_multilingual	0.83	−1.974551462	0.0543
Requests	BERTje	0.78	***Clinical_BERT***	0.84	−2.087584189	***0.0424***
Requests	BERTje	0.78	***BERT_base***	0.83	−2.115004455	***0.0399***
Requests	RobBERT	0.85	BERT_multilingual	0.83	1.135570772	0.262
Requests	RobBERT	0.85	Clinical_BERT	0.84	0.826261555	0.4129
Requests	RobBERT	0.85	BERT_base	0.83	1.084806509	0.2837
Requests	BERT_multilingual	0.83	Clinical_BERT	0.84	−0.261854307	0.7946
Requests	BERT_multilingual	0.83	BERT_base	0.83	−0.111599521	0.9116
Requests	Clinical_BERT	0.84	BERT_base	0.83	0.171520075	0.8646
Reports	BERTje	0.72	RobBERT	0.73	−0.188407237	0.8512
Reports	BERTje	0.72	BERT_multilingual	0.66	1.264072189	0.2113
Reports	BERTje	0.72	Clinical_BERT	0.64	1.59058509	0.1171
Reports	***BERTje***	0.72	BERT_base	0.56	3.136497037	***0.0027***
Reports	RobBERT	0.73	BERT_multilingual	0.66	1.303642356	0.1975
Reports	RobBERT	0.73	Clinical_BERT	0.64	1.606876236	0.1135
Reports	***RobBERT***	0.73	BERT_base	0.56	3.033335703	***0.0036***
Reports	BERT_multilingual	0.66	Clinical_BERT	0.64	0.355254186	0.7237
Reports	BERT_multilingual	0.66	BERT_base	0.56	1.760136321	0.0837
Reports	Clinical_BERT	0.64	BERT_base	0.56	1.356195625	0.1803
	b. Training duration					
Requests	2/3 epochs, 510/765 iterations	0.82	7.8 epochs, 2000 iterations	0.83	−0.588114756	0.5582
Requests	2/3 epochs, 510/765 iterations	0.82	15.7 epochs, 4000 iterations	0.83	−0.900761647	0.3705
Requests	7.8 epochs, 2000 iterations	0.83	15.7 epochs, 4000 iterations	0.83	−0.271829258	0.7865
Reports	2/3 epochs, 510/765 iterations	0.57	***7.8 epochs, 2000 iterations***	0.7	−3.153959736	***0.0021***
Reports	2/3 epochs, 510/765 iterations	0.57	***15.7 epochs, 4000 iterations***	0.71	−3.413500595	***0.0009***
Reports	7.8 epochs, 2000 iterations	0.7	15.7 epochs, 4000 iterations	0.71	−0.37444489	0.7089
	c. Item prevalence					
Requests	***high***	0.89	low	0.81	5.132070191	***<0.0001***
Reports	***high***	0.75	low	0.62	3.393674964	***0.0009***
	d. Dataset					
Requests and reports	***Requests***	0.83	Reports	0.66	8.014726402	***<0.0001***

Table 7. Diagnostic yield per report item. The request items represent the presence of items, irrespective of the number of other items that are also present in the same request. For example, a request with a positive label for "Infiltrate", with or without other categories in the same request, has a 28.1% likelihood of receiving a corresponding report with a positive label for "Infiltrate".

	Report_ Infiltrate	Report_ Decompensation	Report_ Tumor	Report_Pleural_ fluid	Report_Pulmonary_ embolism	Report_ Pneumothorax	Report_ Other	Report_ Groundglass	Report_ COVID19	Report_ Normal
Request_Infiltrate	28.1	4.7	3	16.1	0.3	0.5	76.9	3.3	0.9	20.3
Request_Decompensation	25	14	0.9	27.6	0.2	0.2	86.1	2	0.6	11.6
Request_Tumor	10.9	0.5	19.5	9.5	0.3	0.3	88	3.9	0.2	11.8
Request_Pleural_fluid	22.2	4.8	3.8	38.8	0.3	2.2	85.5	2.4	0.7	11.6
Request_COVID19	33.5	5.2	1.5	13.1	1.3	1	76.2	8.3	2.8	20.4
Request_Pulmonary_embolism	20.6	0.4	4.4	17.3	13.9	0.3	91.6	16.2	6.8	8.4
Request_Pneumothorax	9.2	0.9	1.6	12.5	0	10.9	72.7	1.3	0.1	27.5
Request_Other	11.8	1.5	4.8	8.8	0.3	0.8	73.6	2.7	0.4	25.8

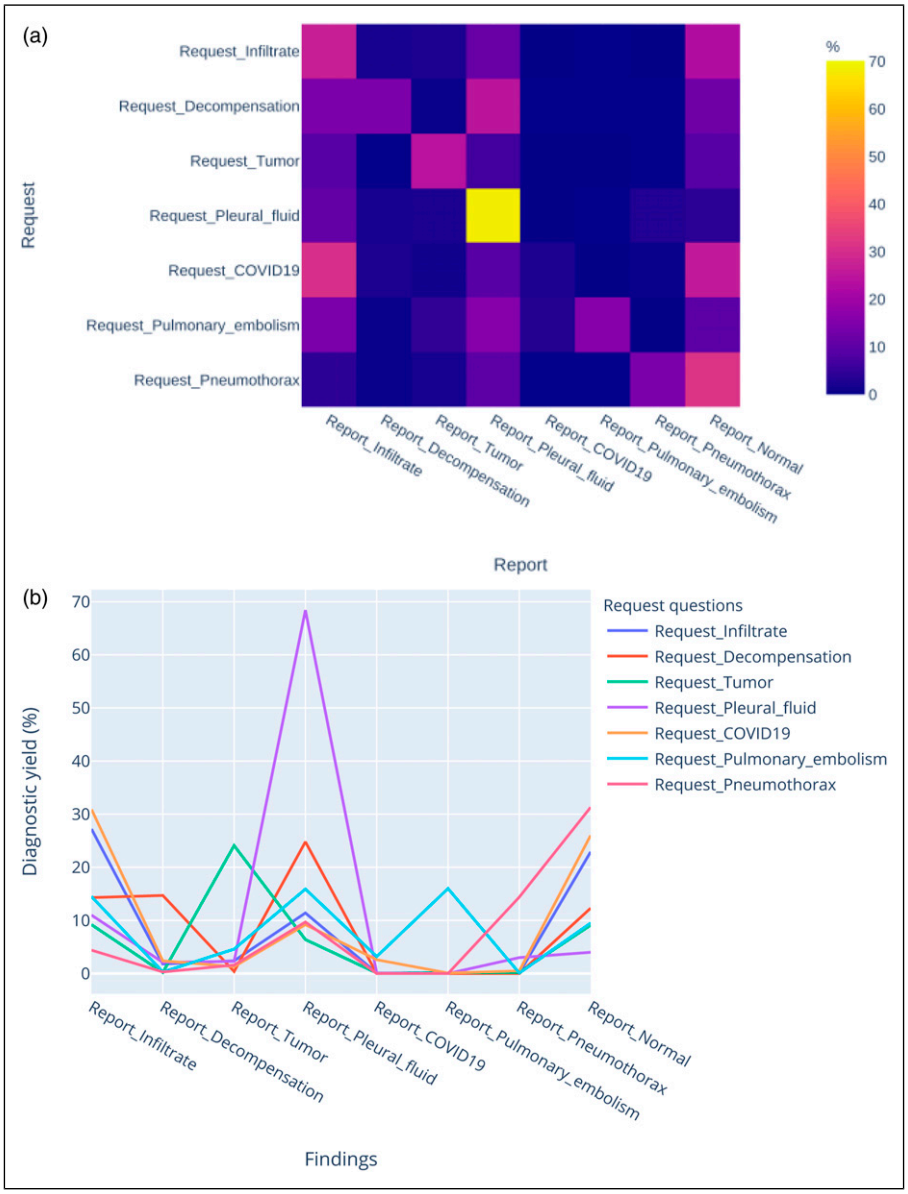

Figure 5. (a) Heatmap of the diagnostic yield of single-request items for different report categories. The colour indicates the percentage of positive findings for a report item. For example, a request mentioning pleural fluid has a 70% chance of receiving a report with a positive result for pleural fluid. (b) Diagnostic yield of different report findings for all single-request items. Some report findings are found only on corresponding requests. For example, pulmonary embolism is only found in examinations with pulmonary embolism on the request. Other imaging findings, like infiltrate, are found in various examinations independent of the request.

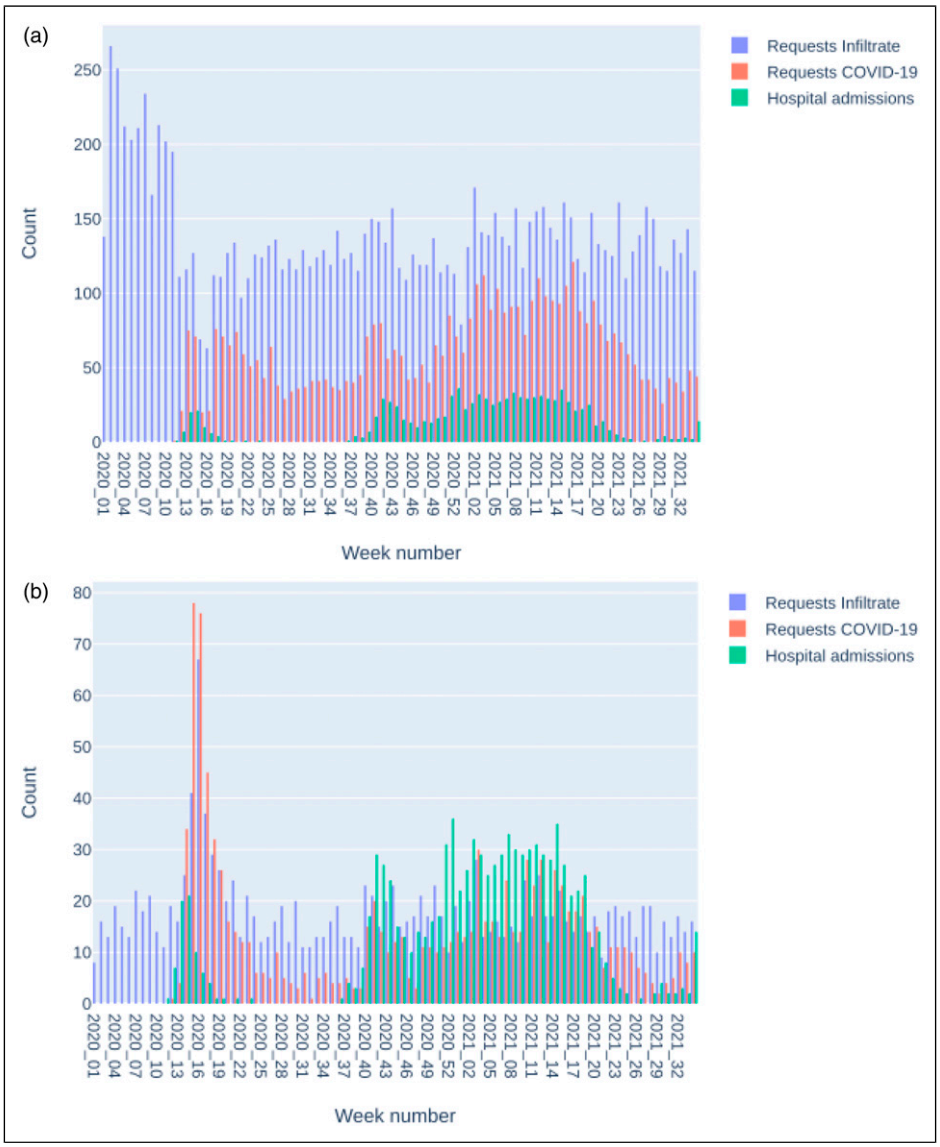

Figure 6. Radiology CR (a.) and CT (b.) requests per week with positive labels in the categories "Infiltrate" or "COVID-19" and hospital admissions per week for COVID-19 patients. The request categories are nonexclusive: each request can have labels in one or more categories.

single person. Thus, it was not possible to assess inter-annotator agreement. To ensure the quality of the dataset, the same person performed an unblinded check for label consistency. External validation of a dataset annotated by multiple radiologists is needed to assess the generalisability of the trained models.

Conclusion

Transformer-based NLP is feasible for the multilabel classification of chest imaging request and report items, even after the fine-tuning needed by pretrained, language-specific models. The developed pipeline makes it possible to combine information from radiology requests and reports on a large scale to assess radiology utilization and diagnostic yield. Diagnostic yield in chest imaging varies with the information in the requests; therefore, the inclusion of NLP analyses of requests is recommended for quality control of and research into chest imaging.

Highlights

- Transformer-based natural language processing is feasible for the multilabel classification of radiology requests and reports.
- Training and testing on a dataset containing 2256 requests and reports demonstrated good results.
- Among five transformer models and one LSTM model, the RobBERT model surpassed the others and was used for the multilabel classification of 40,873 radiology requests and reports.
- *Di*agnostic yield in chest imaging varies with the information in the requests.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Allard W Olthof https://orcid.org/0000-0003-0809-9722
Peter MA van Ooijen https://orcid.org/0000-0002-8995-1210

Supplemental Material

Supplemental material for this article is available online.

References

1. Eberhardt SC and Heilbrun ME. Radiology report value equation. *Radiographics* 2018; 38(5): 1888–1896, DOI: 10.1148/rg.2018180133.
2. Goldberg-Stein S and Chernyak V. Adding Value in Radiology Reporting. *J Am Coll Radiol* 2019; 16: 1292–1298, DOI: 10.1016/j.jacr.2019.05.042.
3. Mehta N and Pandit A. Concurrence of big data analytics and healthcare: A systematic review. *Int J Med Inform* 2018; 114: 57–65, DOI: 10.1016/j.ijmedinf.2018.03.013.
4. Jones S, Cournane S, Sheehy N, et al. A Business Analytics Software Tool for Monitoring and Predicting Radiology Throughput Performance. *J Digit Imaging* 2016; 29(6): 645–653, DOI: 10.1007/s10278-016-9871-3.

5. Shailam R, Botwin A, Stout M, et al. Real-Time Electronic Dashboard Technology and Its Use to Improve Pediatric Radiology Workflow. *Curr Probl Diagn Radiol* 2018; 47(1): 3–5, DOI: 10.1067/j.cpradiol.2017.03.002.
6. Rehani MM, Melick ER, Alvi RM, et al. Patients undergoing recurrent CT exams: assessment of patients with non-malignant diseases, reasons for imaging and imaging appropriateness. *Eur Radiol* 2020; 30(4): 1839–1846, DOI: 10.1007/s00330-019-06551-8.
7. Fernandez M and Craig S. Appropriateness of adult plain abdominal radiograph requesting in a regional Emergency Department. *J Med Imaging Radiat Oncol* 2019; 63(2): 175–182, DOI: 10.1111/1754-9485.12847.
8. Jarvik JG, Meier EN, James KT, et al. The Effect of Including Benchmark Prevalence Data of Common Imaging Findings in Spine Image Reports on Health Care Utilization Among Adults Undergoing Spine Imaging: A Stepped-Wedge Randomized Clinical Trial. *JAMA Netw Open* 2020; 3(9): e2015713, DOI: 10.1001/jamanetworkopen.2020.15713.
9. Common J, Ramonas M and Alabousi A. The Diagnostic Yield of CT Urography in the Workup of Hematuria With Negative Cystoscopy. *Can Assoc Radiol J* 2021; 72(4): 728–735, DOI: 10.1177/0846537120933952.
10. Murray M and Costa AF. Appropriateness of Abdominal Aortic Aneurysm Screening With Ultrasound: Potential Cost Savings With Guideline Adherence and Review of Prior Imaging. *Can Assoc Radiol J* 2021; 72(3): 398–403, DOI: 10.1177/0846537120920866.
11. Cho MS, Roh JH, Park H, et al. Practice Pattern, Diagnostic Yield, and Long-Term Prognostic Impact of Coronary Computed Tomographic Angiography. *J Am Heart Assoc* 2020; 9(18): e016620, DOI: 10.1161/JAHA.120.016620.
12. Guarnizo A, Farah K, Lelli DA, et al. Limited usefulness of routine head and neck CT angiogram in the imaging assessment of dizziness in the emergency department. *Neuroradiol J* 2021; 34(4): 1971400920988665, DOI: 10.1177/1971400920988665.
13. Richardson S, Lucas E, Cohen SL, et al. Predictors of Overtesting in Pulmonary Embolism Diagnosis. *Acad Radiol* 2020; 27(3): 404–408, DOI: 10.1016/j.acra.2019.04.018.
14. Carrell DS, Halgrim S, Tran D-T, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol* 2014; 179(6): 749–758, DOI: 10.1093/aje/kwt441.
15. Sorin V, Barash Y, Konen E, et al. Deep Learning for Natural Language Processing in Radiology—Fundamentals and a Systematic Review. *J Am Coll Radiol* 2020; 17(5): 639–648, DOI: 10.1016/j.jacr.2019.12.026.
16. Kang SK, Garry K, Chung R, et al. Natural Language Processing for Identification of Incidental Pulmonary Nodules in Radiology Reports. *J Am Coll Radiol* 2019; 16(11), DOI: 10.1016/j.jacr.2019.04.026.
17. Dublin S, Baldwin E, Walker RL, et al. Natural Language Processing to identify pneumonia from radiology reports. *Pharmacoepidemiol Drug Saf* 2013; 22(8): 834–841, DOI: 10.1002/pds.3418.
18. Huesch MD, Cherian R, Labib S, et al. Evaluating Report Text Variation and Informativeness: Natural Language Processing of CT Chest Imaging for Pulmonary Embolism. *J Am Coll Radiol* 2018; 15(3 Pt B): 554–562, DOI: 10.1016/j.jacr.2017.12.017.
19. Chen MC, Ball RL, Yang L, et al. Deep Learning to Classify Radiology Free-Text Reports. *Radiology* 2018; 286(3): 845–852, DOI: 10.1148/radiol.2017171115.
20. Moradi M, Dorffner G and Samwald M. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Comput Methods Programs Biomed* 2020; 184: 105117, DOI: 10.1016/j.cmpb.2019.105117.
21. Zhang X, Zhang Y, Zhang Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int J Med Inform* 2019; 132, DOI: 10.1016/j.ijmedinf.2019.103985.

22. Wei Q, Ji Z, Si Y, et al. Relation Extraction from Clinical Narratives Using Pre-trained Language Models. *AMIA: Annu Symp Proceedings AMIA Symp* 2019; 2019: 1236–1245.
23. Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* 2018; 1810: 04805v2.
24. Kuling G, Curpen B and Martel AL. BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports. *J Imaging* 2022; 8(5), DOI: 10.3390/JIMAGING8050131.
25. Datta S and Roberts K. A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (ACL), 2020, pp. 50–55, DOI: 10.18653/v1/2020.splu-1.6.
26. Nakamura Y, Hanaoka S, Nomura Y, et al. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. *BMC Med Inform Decis Mak* 2021; 21(1): 262, DOI: 10.1186/s12911-021-01623-6.
27. Niehues SM, Adams LC, Gaudin RA, et al. Deep-Learning-Based Diagnosis of Bedside Chest X-ray in Intensive Care and Emergency Medicine. *Invest Radiol* 2021; 56(8): 525–534, DOI: 10.1097/RLI.0000000000000771.
28. Alfarghaly O, Khaled R, Elkorany A, et al. Automated radiology report generation using conditioned transformers. *Inform Med Unlocked* 2021; 24: 100557, DOI: 10.1016/J.IMU.2021.100557.
29. Olthof AW, van Ooijen PMA and Cornelissen LJ. Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. *J Med Syst* 2021; 45(10), DOI: 10.1007/S10916-021-01761-4.
30. de Vries W, van Cranenburgh A, Bisazza A, et al. BERTje: A Dutch BERT Model. *arXiv* 2019; 1912: 09582.
31. Delobelle P, Winters T and Berendt B. RobBERT: A Dutch RoBERTa-based language model. In: *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. Association for Computational Linguistics (ACL), 2020, pp. 3255–3265, DOI: 10.18653/v1/2020.findings-emnlp.292.
32. Pires T, Schlinger E and Garrette D. How multilingual is Multilingual BERT? ACL 2019. In: 57th Annu Meet Assoc Comput Linguist Proc Conf, 2019, pp. 4996–5001, DOI: 10.18653/v1/p19-1493.
33. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. *arXiv* 2019; 1904: 03323.
34. Sechidis K, Tsoumakas G and Vlahavas I. On the stratification of multi-label data. *Mach Learn Knowl Discov Databases* 2011; 6913: 145–158.
35. Szymański P and Kajdanowicz T. A Network Perspective on Stratification of Multi-Label Data. In: Torgo L, Krawczyk B, Branco P and Moniz N (eds). Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications. ECML-PKDD. Skopje, Macedonia: PMLRProceedings of Machine Learning Research, 2017, pp. 22–35. vol. 74.
36. Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-Art Natural Language Processing. *arXiv* 2020; 1910: 03771v5, DOI: 10.18653/v1/2020.emnlp-demos.6.
37. National Institute for Public Health and the Environment. *Covid-19 ziekenhuisopnames (volgens NICE registratie) per gemeente per ziekenhuisopnamedatum en meldingsdatum | Data overheid*, 2022. [Internet]. [cited Jan 12]. Available from: https://data.overheid.nl/dataset/14845-covid-19-ziekenhuisopnames–volgens-nice-registratie–per-gemeente-per-ziekenhuisopnamedatum-e#panel-resources
38. Sorantin E, Grasser MG, Hemmelmayr A, et al. The augmented radiologist: artificial intelligence in the practice of radiology. *Pediatr Radiol* 2021; 1–13, DOI: 10.1007/s00247-021-05177-7.
39. Wood DA, Kafiabadi S, Al Busaidi A, et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. *Eur Radiol* 2022; 32(1): 725–736, DOI: 10.1007/s00330-021-08132-0.

40. Bressem KK, Adams LC, Gaudin RA, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* 2020; 36(21): 5255–5261, DOI: 10.1093/bioinformatics/btaa668.
41. Venturelli F, Ottone M, Pignatti F, et al. Using text analysis software to identify determinants of inappropriate clinical question reporting and diagnostic procedure referrals in Reggio Emilia, Italy. *BMC Health Serv Res* 2021; 21(1), DOI: 10.1186/s12913-021-06093-0.
42. Viau JA, Chaudry H, Hannigan A, et al. The Yield of Computed Tomography of the Head Among Patients Presenting With Syncope: A Systematic Review. *Acad Emerg Med* 2019; 26(5): 479–490, DOI: 10.1111/acem.13568.
43. Pons E, Foks KA, Dippel DWJ, et al. Impact of guidelines for the management of minor head injury on the utilization and diagnostic yield of CT over two decades, using natural language processing in a large dataset. *Eur Radiol* 2019; 29(5): 2632–2640, DOI: 10.1007/s00330-018-5954-5.
44. Annarumma M, Withey SJ, Bakewell RJ, et al. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* 2019; 291(1): 196–202, DOI: 10.1148/radiol.2018180921.
45. Short RG, Bralich J, Bogaty D, et al. Comprehensive Word-Level Classification of Screening Mammography Reports Using a Neural Network Sequence Labeling Approach. *J Digit Imaging* 2019; 32(5): 685–692, DOI: 10.1007/s10278-018-0141-4.
46. Lu H, Ehwerhemuepha L and Rakovski C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med Res Methodol* 2022; 22(1): 1–12, DOI: 10.1186/S12874-022-01665-Y.
47. Carrington AM, Fieguth PW, Qazi H, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med Inform Decis Mak* 2020; 20(1): 1–12, DOI: 10.1186/S12911-019-1014-6/FIGURES/9.