Inaugural – Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Klaus Kades, M.Sc.

aus Werneck, Germany

Tag der mündlichen Prüfung: ______________________________

Current Challenges in the Application of Algorithms in Multi-institutional Clinical Settings

Supervisor: Prof. Dr. Klaus H. Maier-Hein

Abstract

The Coronavirus disease pandemic has highlighted the importance of artificial intelligence in multi-institutional clinical settings. Particularly in situations where the healthcare system is overloaded, and a lot of data is generated, artificial intelligence has great potential to provide automated solutions and to unlock the untapped potential of acquired data. This includes the areas of care, logistics, and diagnosis. For example, automated decision support applications could tremendously help physicians in their daily clinical routine. Especially in radiology and oncology, the exponential growth of imaging data, triggered by a rising number of patients, leads to a permanent overload of the healthcare system, making the use of artificial intelligence inevitable. However, the efficient and advantageous application of artificial intelligence in multi-institutional clinical settings faces several challenges, such as accountability and regulation hurdles, implementation challenges, and fairness considerations. This work focuses on the implementation challenges, which include the following questions: How to ensure well-curated and standardized data, how do algorithms from other domains perform on multi-institutional medical datasets, and how to train more robust and generalizable models? Also, questions of how to interpret results and whether there exist correlations between the performance of the models and the characteristics of the underlying data are part of the work. Therefore, besides presenting a technical solution for manual data annotation and tagging for medical images, a real-world federated learning implementation for image segmentation is introduced. Experiments on a multi-institutional prostate magnetic resonance imaging dataset showcase that models trained by federated learning can achieve similar performance to training on pooled data. Furthermore, Natural Language Processing algorithms with the tasks of semantic textual similarity, text classification, and text summarization are applied to multi-institutional, structured and free-text, oncology reports. The results show that performance gains are achieved by customizing state-of-the-art algorithms to the peculiarities of the medical datasets, such as the occurrence of medications, numbers, or dates. In addition, performance influences are observed depending on the characteristics of the data, such as lexical complexity. The generated results,

human baselines, and retrospective human evaluations demonstrate that artificial intelligence algorithms have great potential for use in clinical settings. However, due to the difficulty of processing domain-specific data, there still exists a performance gap between the algorithms and the medical experts. In the future, it is therefore essential to improve the interoperability and standardization of data, as well as to continue working on algorithms to perform well on medical, possibly, domain-shifted data from multiple clinical centers.

Zusammenfassung

Die Coronavirus-Pandemie hat die Bedeutung von künstlicher Intelligenz in multizentrischen klinischen Settings besonders deutlich gemacht. Vor allem in Situationen, in denen das Gesundheitssystem überlastet ist und gleichzeitig viele Daten generiert werden, kann künstliche Intelligenz automatisierte Lösungen anbieten, um das hohe Potenzial erfasster Daten besser zu nutzen. Dies trifft insbesondere in den Bereichen Pflege, Logistik und Diagnose zu. So könnten beispielsweise automatisierte Anwendungen Ärzte in ihrer täglichen klinischen Routine bei Entscheidungen enorm unterstützen. Gerade in der Radiologie und Onkologie führt das exponentielle Wachstum der auszuwertenden Bilddaten, unter anderem ausgelöst durch eine steigende Zahl von Patienten, zu einer permanenten Überlastung des Gesundheitssystems, was den Einsatz von künstlicher Intelligenz unumgänglich macht. Die effiziente und vorteilbringende Anwendung von künstlicher Intelligenz in klinischen Settings mit mehreren Institutionen steht jedoch vor verschiedenen Herausforderungen, wie zum Beispiel Hürden bei Verantwortlichkeiten und Regulierungen, Implementierungsproblemen und Fairnessüberlegungen. Diese Arbeit konzentriert sich auf die Herausforderungen bei der Implementierung, zu denen die folgenden Fragestellungen gehören: Wie können gut kuratierte und standardisierte Daten erstellt werden? Wie schneiden Algorithmen aus anderen Domänen angewandt auf medizinische Datensätze verschiedener Institutionen ab? Und wie können robuste und generalisierte Modelle trainiert werden? Weiterhin wird diskutiert, wie die Ergebnisse zu interpretieren sind und ob es Korrelationen zwischen der Performance der Modelle und den Eigenschaften der zugrunde liegenden Daten gibt. Daher wird in dieser Arbeit nicht nur eine technische Lösung für die manuelle Datenannotation und das Tagging von medizinischen Bildern vorgestellt, sondern auch eine praxisbezogene Implementierung von föderiertem Lernen für die Bildsegmentierung präsentiert. Experimente auf einem Prostata-Datensatz verschiedener Institutionen zeigen, dass Modelle, die durch föderiertes Lernen trainiert werden, ähnliche Ergebnisse erzielen können wie durch das Training auf zusammengeführten Daten. Darüber hinaus werden Algorithmen der natürlichen Sprachverarbeitung auf strukturierte und Freitext-Onkologie Befunde mehrerer Institutionen angewendet.

Hierbei werden vor allem die Themenbereiche der semantischen Ähnlichkeit zwischen Texten, sowie der Klassifizierung und der Zusammenfassung von Texten behandelt. Die Ergebnisse zeigen, dass die state-of-the-art Algorithmen eine verbesserte Performance erreichen können, indem diese an die Besonderheiten der medizinischen Datensätze anpasst werden. Dies betrifft zum Beispiel das Vorkommen von Medikamenten, Zahlen oder Daten in medizinischen Texten. Darüber hinaus werden Leistungsunterschiede in Abhängigkeit von den Eigenschaften der Daten wie der lexikalischen Komplexität beobachtet. Die generierten Ergebnisse, die Baselines der Annotatoren und die retrospektiven Bewertungen von Annotatoren zeigen, dass Algorithmen der künstlichen Intelligenz großes Potenzial für den Einsatz im klinischen Umfeld haben. Allerdings besteht aufgrund der erschwerten Verarbeitung domänenspezifischer Daten immer noch eine Lücke zwischen der Performance von Algorithmen und medizinischen Experten. In Zukunft ist es demnach wichtig, die Interoperabilität und die Standardisierung von Daten zu verbessern und weiterhin an Algorithmen zu arbeiten, die auf medizinischen, möglicherweise multizentrischen Daten aus mehreren Kliniken gut funktionieren.

Acknowledgements

Writing a doctoral thesis is an enormous commitment over multiple years, with many ups and downs and important decisions along the way. Especially because of this commitment, the colleagues, the supervisor, and the social environment are of very high importance. This is why I would like to express my sincere gratitude to all those who have supported, guided, and accompanied me along the way.

A special thanks goes to my supervisor, Klaus Maier-Hein, for giving me the incredible possibility to work over multiple years on different, always very interesting projects. I want to thank him especially for the very honest and open conversations, which always gave me faith to continue, and for the awesome working environment that he creates within his group. Furthermore, I want to thank my medical supervisor, Jens Kleesiek. He was a second pillar on whom I could rely in times of doubt, who kept me motivated with his visions, and who, as my medical cooperation partner, was a driving factor for the contents of my thesis.

I would like to thank everyone who guided and mentored me throughout my thesis. This includes but is not limited to Dr. Peter Neher, Dr. Ralf Floca, Dr. Marco Nolden, Prof. Michael Strube, Dr. Fabian Isensee, and Dr. Paul Jäger. Furthermore, I want to thank Verena Schneier, Jasmin Metzger, Dr. Marco Nolden and again Prof. Jens Kleesiek for the great times working on the Joint Imaging Platform (JIP) project, which was successful despite the great challenge thanks to the extraordinary team. Furthermore, I want to thank Dr. Tobias Penzkofer and Dr. Andreas Bucher for the great experiences and insights I could gather during my time working on the Radiological Cooperative Network (RACOON) project. Also, I would like to thank Dr. Matthias Fink for being an incredible medical partner as well as Prof. Tim Weber for his collaborations. Moreover, I would like to thank everyone who provided me with text annotations for the medical reports, including Jens Kleesiek, Matthias Fink, Peter Full, Arved Bishoff, Martin Moll, Merle Schnell, Maike Küchler, Melina Fritz, Julia Jaworek, and Genoveva Wolf.

I am indescribably grateful for having the chance to be part of a very strong and diverse team, which I want to thank for all the nice discussions, experiences, and moments I had throughout my time as a doctoral student. Especially, I want to thank Jonas

Scherer for being the best teammate one could think of, for trusting and believing in me, and for giving me the freedom I sometimes asked for. I would like to thank my colleagues Jan Sellner, Max Zenk and Gregor Köhler, who have accompanied me throughout the years, listened to my problems, and have always given me valuable input. I would like to thank my two Master students, Siting Liang and Marius Kempf, for their incredible work and for giving me the possibility to learn skills as a supervisor. I want to thank the whole Kaapana team, for carrying on the passion that Jonas and I put into the project. Furthermore, I want to thank various colleagues who have supported me as friends over the years and who helped me with the actual writing of my thesis, including Lisa Kausch, David Zimmerer, Hanno Gao, Stefan Denner, Markus Bujotzek, Philipp Schader, Melanie Schellenberg, Lukas Kades and again Jonas Scherer, Max Zenk, Jan Sellner, Matthias Fink and Siting Liang. If I missed someone in this section, that is on me, but know that I appreciate every little piece of help.

Finally, I would like to express my deepest gratitude to my friends and especially to my family. They were the backbone of everything I have done and supported me whenever I needed them. From the bottom of my heart, thank you.

Contents

Abstract **i**

Acknowledgements **v**

Contents **vii**

Acronyms **xiii**

List of figures **xix**

List of tables **xxii**

1 Introduction **1**

1.1 The bigger picture 1

1.2 Status quo 4

1.2.1 The data landscape in the clinics 4

1.2.2 The algorithmic landscape in the clinics 6

Rule-based systems, machine learning, neural networks, and deep neural networks 6

Natural Language Processing algorithms 7

Medical image computing 8

1.2.3 Real-world initiatives between multiple clinics 9

1.3 Open challenges for the application of algorithms in multi-institutional clinical settings 12

1.3.1 Challenges 12

1.3.2 Solution approaches 13

Standardization and curation 13

Clinical adaptation 14

Validation of artificial intelligence methods 15

Trustworthiness, explainability and interpretability 16

1.4 Contributions and outline 18
1.4.1 Semantic modeling for semantic textual similarity 18
1.4.2 Semi-structured data analysis for text summarization 19
1.4.3 Assessing distributional shifts for text classification 20
1.4.4 Manual data annotation and tagging for medical images 21
1.4.5 Real-world federated learning for the task of image segmentation 22

2 Related work 25
2.1 Semantic modeling for semantic textual similarity 26
2.2 Semi-structured data analysis for text summarization 27
2.3 Assessing distributional shifts for text classification 29
2.4 Manual data annotation and tagging for medical images 31
2.5 Real-world federated learning for the task of image segmentation . . . 32

3 Materials 35
3.1 Kaapana as an open-source toolkit for provisioning medical imaging platforms 35
3.1.1 The technological foundation of Kaapana 35
The IaaS and CaaS layer 36
The PaaS layer - SYSTEM and MONITORING units 37
The SaaS layer - BASE, STORE and META units 37
The FaaS layer - FLOW unit 38
3.1.2 Deployment and use cases of Kaapana 39
Kaapana deployment within the DKTK and RACOON 39
nnU-Net training as use case 40
3.2 Multi-institutional datasets 44
3.2.1 Clinical Semantic Textual Similarity dataset 44
3.2.2 German radiology reports 45
3.2.3 The multi-site prostate MRI segmentation dataset 47

4 Methods 53
4.1 Semantic modeling for semantic textual similarity 53
4.1.1 Overview 53
Baseline 53
Cluster analysis 54
4.1.2 Approaches 54
Enhancing BERT with features based on similarity measures . . 54
M-Heads 55
Medication graph 57
4.2 Semi-structured data analysis for text summarization 62

4.2.1 Approaches 62
Abstractive text summarization baseline 62
Baseline and extraction 63
Baseline and pointer 65
Baseline combining extraction and pointer 66
Greedy search, N-gram blocking and repetition penalty 66
4.2.2 Human evaluation 67
4.3 Assessing distributional shifts for text classification 69
4.3.1 Analysis of report corpora 69
Medical characteristics of the report corpora 69
Complexity analysis of report corpora 69
4.3.2 Reference annotations 70
Automated reference annotation from SORs 70
Manuel reference annotation of FTORs 71
4.3.3 Text classification baselines 71
Human baseline on FTORs 71
Text classification algorithm 71
4.3.4 Interpretability and explainability of the models 73
Highlighting important tokens 73
Vizualization of the feature vectors 74
Calibration of the model's predictions 74
4.4 Manual data annotation and tagging for medical images 76
4.4.1 The used open-source software toolkits 76
4.4.2 Combining the different toolkits 77
4.5 Real-world federated learning for the task of image segmentation . . . 80
4.5.1 Backend and user interface 80
4.5.2 Local workflow adjustments 81
4.5.3 The nnU-Net in federated settings 82

5 Experiments and results **83**
5.1 Semantic modeling for semantic textual similarity 83
5.1.1 Experiment setup 83
Preprocessing 83
Feature sets and voting regressor for the EnhancedBERT 84
Medication graph training 85
Dataset analysis 85
Evaluation runs 85
5.1.2 Results 87
Dataset evaluation 87
Evaluation runs 88

5.2 Semi-structured data analysis for text summarization . . . 91
5.2.1 Experiment setup . . . 91
Preprocessing steps . . . 91
Evaluation runs . . . 91
5.2.2 Results . . . 92
Human evaluation . . . 93
5.3 Assessing distributional shifts for text classification . . . 100
5.3.1 Experiment setup . . . 100
The report corpora . . . 100
Human annotations . . . 102
NLP models . . . 102
Combining human- and machine-generated annotations . . . 103
Evaluation and statistical tests . . . 103
5.3.2 Results . . . 105
The report corpora . . . 105
Human annotations . . . 107
NLP models runs . . . 111
Combining human- and machine-generated annotations . . . 111
Evaluation and statistical tests . . . 111
5.4 Manual data annotation and tagging for medical images . . . 132
5.4.1 Experiment setup . . . 132
5.4.2 Results . . . 132
5.5 Real-world federated learning for the task of image segmentation . . . 135
5.5.1 Experiment setup . . . 135
The experiments . . . 135
Training of the nnU-Net . . . 136
5.5.2 Results . . . 137
Feasibility of training the nnU-Net in a federated way . . . 137
Experiment results . . . 140

6 Discussion 143
6.1 Semantic modeling for semantic textual similarity . . . 143
6.1.1 Approach 1: Voting regression . . . 143
6.1.2 Approach 2: M-Heads . . . 143
6.1.3 Approach 3: Medication graph . . . 144
6.1.4 Limitations . . . 145
6.2 Semi-structured data analysis for text summarization . . . 146
6.2.1 Automatic evaluation . . . 146
6.2.2 Expert evaluation . . . 146
6.2.3 Limitations . . . 147

6.3 Assessing distributional shifts for text classification 148
6.3.1 The challenges of the dataset 148
6.3.2 The difficulty of a correct tumor response category assignment . 148
6.3.3 The human and machine baselines 149
6.3.4 The importance of confidences 150
6.3.5 Combining human- and machine-generated annotations 151
6.3.6 Correlations between dataset characteristics and performances . 151
6.3.7 Limitations 152
6.4 Manual data annotation and tagging for medical images 154
6.4.1 The implementation 154
6.4.2 Limitations 154
6.5 Real-world federated learning for the task of image segmentation . . . 155
6.5.1 The implementation 155
6.5.2 Experimental results 155
6.5.3 Limitations 156

7 Conclusion 159

A Own contributions and publications 165
A.1 Own contributions 166
A.2 Own publications 168

B Appendix 171
B.1 Annotation guides and screenshots 171
B.1.1 Annotation guide for TRC assignment of the FTORs 171
B.1.2 Annotation guide for expert evaluation on generated conclusions 172
B.1.3 Screenhots of the annotation software 173
B.2 Tables 175

Bibliography 179

Acronyms

AE Application Entity

AI artificial intelligence

ANOVA Analysis Of Variance

API application programming interface

ASD Average Surface Distance

AUC Area under the receiver operating characteristic curve

BERT Bidirectional Encoder Representation from Transformers

BMBF German Federal Ministry of Education and Research

BN Bayesian Networks

BOW Bag-of-Words

BraTS Brain Tumor Segmentation

cf. compare

CG central gland

CI Confidence Interval

CIS Clinical Information System

NVIDIA Clara Computational platform to build manage and deploy intelligent medical imaging workflows and instruments

Clinical STS Clinical Semantic Textual Similarity

CLS token Classifier token

CNN Convolutional Neural Network

COVID-19 Coronavirus disease

CR complete response

CRF Conditional Random Fields

CT computed tomography

CUP cancer of unknown primary

CV Computer Vision

DAG Directed Acyclic Graph

dcmqi DICOM for Quantitative Imaging

DICE Sørensen-Dice coefficient

DICOM Digital Imaging and Communications in Medicine

DICOM SEG DICOM Segmentation Objects

DKFZ German Cancer Research Center

DKTK German Cancer Consortium

DL deep learning

DNN deep neural network

DT Decision Trees

ECE Expected Calibration Error

e.g. exempli gratia

EHR Electronic health record

FAIR Findable, Accessible, Interoperable and Reusable

FHIR Fast Healthcare Interoperability Resources

FL Federated Learning

FTOR free-text-oncology report

FTOR-DKFZ free-text-oncology reports from the German Cancer Research Center

FTORT-TKH free-text-oncology reports from the Heidelberg Thoracic Clinic

GloVe Global Vectors for Word Representation

GLUE General Language Understanding Evaluation

GPU Graphics Processing Unit

GPT-2 Generative Pre-trained Transformer 2

HIS Hospital Information System

HL7 Health Level 7

HTML HyperText Markup Language

HTTPS Hypertext Transfer Protocol Secure

i2b2 Informatics for Integrating Biology & the Bedside

ICD International Classification of Diseases

i.e. id est

IT information technology

iframe inline frame

JIP Joint Imaging Platform

Kaapana Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis

KNN K-nearest neighbors

LDA Latent Dirichlet Allocation

Linear-SVC Linear Support Vector Classifier

LIS Laboratory Information System

LR Logistic Regression

LSA Latent Semantic Analysis

LSTM Long short-term memory

MiA micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.)

MIC medical image computing

MinIO High Performance Object Storage

MITK Medical Imaging Interaction Toolkit

ML machine learning

MNB Multinomial Naïve Bayes

MONAI Medical Open Network for Artificial Intelligence

MRI magnetic resonance imaging

MR magnetic resonance

MSE Mean Squared Error

N2C2 National NLP Clinical Challenges

NIfTI Neuroimaging Informatics Technology Initiative

NLP Natural Language Processing

NN Neural network

nnU-Net no new net U-Net

Non-IID not independent and identically distributed

Nrrd "nearly raw raster data"

NUM Network of University Medicine

OCI Open Container Initiative

OCT Optical Coherence Tomography

OHIF Open Health Imaging Foundation

OHNLP Open Health Natural Language Processing

OIDC OpenID Connect

OS operating system

PACS Picture Archiving and Communication System

PCC Pearson correlation coefficient

PCA Principal Component Analysis

PD progressive disease

PET positron emission tomography

PR partial response

PriMIA Privacy-preserving Medical Image Analysis

PZ peripheral zone

RACOON Radiological Cooperative Network

RECIST Response evaluation criteria in solid tumors

regex Regular expression

RF Random Forests

RIS Radiology Information System

RNN Recurrent Neural Network

ROC receiver operating characteristic

RT radiology technologist

SD stable disease

SEG Segmentation Objects

ROUGE Recall-Oriented Understudy for Gisting Evaluation

SEP token Separator token

seq2seq sequence to sequence

SMART Substitutable Medical Applications, Reusable Technologies

SPA single-page application

SOR Structured Oncology Report

SSO single sign-on

SSL Secure Sockets Layer

STD standard deviation

STS semantic textual similarity

SVM Support Vector Machines

SVR Support Vector Regressor

TAR tape archive

TCIA the Cancer Imaging Archive

TF-IDF term frequency-inverse document frequency

TKH Heidelberg Thoracic Clinic

TRC tumor response category

t-SNE t-distributed stochastic neighbor embedding

Tukey's HSD Tukey's honestly significant difference

UI user interface

UKHD University Hospital Heidelberg

UMAP Uniform Manifold Approximation and Projection

URL Uniform Resource Locator

US ultrasound

List of Figures

1.1 Locations of the cooperation partners within the Joint Imaging Platform (JIP) initiative. 10
1.2 Locations of the cooperation partners within the Radiological Cooperative Network (RACOON). 11
1.3 Challenges and potential solution approaches for the application of algorithms in multi-institutional clinical settings. 17
1.4 Overview of the five main contributions for the application of algorithms in multi-institutional clinical settings 18

3.1 Illustration of the private cloud computing setup of Kaapana consisting of the different service layers. 36
3.2 Setup of the JIP within a clinical center. 40
3.3 Infrastructure stack as it is deployed within RACOON. 41
3.4 Detailed view of the processing pipeline for the nnU-Net training DAG and its corresponding operators. 42
3.5 Detailed view of the processing pipeline for the nnU-Net ensemble DAG and its corresponding operators. 43
3.6 Process of the creation of the dataset for the 2019 N2C2/OHNLP Clinical STS track 44
3.7 Examples of the sentence pairs from the BioCreative/OHNLP challenge 2018. 49
3.8 Oncologic assessment in the clinical routine using structured oncology reporting. 50
3.9 Samples including the prostate ground annotation from the multi-site prostate MRI segmentation dataset. 51

4.1 Illustration of the processing pipelines for the different STS approaches. 55
4.2 Illustration of the medication graph, modeling similarities between agent pairs. 60

4.3 Architecture of the BERT2BERT model. 63
4.4 Architecture of the BERT2BERT model with the extraction loss. 64
4.5 Architecture of the BERT2BERT model incorporating the pointer network. 68
4.6 Illustration of the annotation process within the Kaapana platform. . . . 78

5.1 Distribution of labels and number of words for the training and test dataset. 86
5.2 Visualizations of the training and test dataset using t-SNE projected InferSent embeddings. 87
5.3 Visualizations of the clusters, emerging from the k-means clustering, using t-SNE projected InferSent embeddings. 89
5.4 Illustration of the difference between the reference annotations and the BERT predicted scores along with the number of sentence pairs per cluster. 90
5.5 Scores assigned during the human evaluation for the reference summaries and the summaries generated by the different presented methods. 95
5.6 Rank distribution resulting from the human evaluation for all samples and, respectively, for each of the three criteria: non-oncological and oncological correctness and comprehensibility. 96
5.7 Visualization of the mean rank resulting from the human evaluation across the three criteria and the four TRCs. 97
5.8 Flowchart illustrating the study design and the creation of the training and test sets. 101
5.9 Visualization of the hyperparameter importance and the weighted F1 score as a function of the hyperparameter. 104
5.10 Complexity measures of the three test datasets in comparison to those of three open-source German datasets. 106
5.11 Confusion matrices resulting from the export annotation of the FTOR-DKFZ and FTORT-TKH. 107
5.12 Heatmap, reporting the inter-annotator agreement using the Cohen's kappa between the seven human annotators (two radiologists, two students and three RTs) 108
5.13 Average confidence of the seven annotators along with the STDs on the two FTORs. 108
5.14 Accuracies within binned confidence intervals of the seven annotators on the two FTORs. 109
5.15 Average confidence of the four NLP models along with the STDs on the two FTORs and the training and test splits of the SORs. 110
5.16 Accuracies within binned confidence intervals of the four NLP models on the two FTORs and the training and test splits of the SORs. 110
5.17 Performance differences when ensembling or micro-averaging annotations over different subgroups. 112

5.18 ROC curves for the BERT model and the Linear-SVC on the three test datasets along with the operating points for the three annotator groups on the two FTORs. 113
5.19 Rank distributions of the seven human annotators and the four NLP models on the FTOR-DKFZ and FTORT-TKH. 121
5.20 Exemplary longitudinal representation of the oncologic course for six different patients 122
5.21 Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of grouped confidence intervals of the annotators and two NLP models. 124
5.22 Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the lexical complexity for the FTOR-DKFZ dataset. 125
5.23 Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the lexical complexity for the FTORT-TKH dataset. 126
5.24 Confidences of the BERT model, the Linear-SVC and the three annotators groups as a function of the concordance of oncologic and nononcologic findings described in the FTORT-TKH. 127
5.25 Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the concordance of oncologic and nononcologic findings described in the FTORT-TKH. 128
5.26 Visualization of token importance on a sample findings section of a FTORT-TKH for the four NLP models. 129
5.27 Visualization of token importance on a sample findings section of a FTORT-TKH for the BERT models trained on the $k = 5$ training folds. . . 130
5.28 Visualizations of the CLS token embeddings using UMAP projections on the three test datasets. 131
5.29 Screenshot of the tagging process within Kaapana. 133
5.30 Overview of the federated nnU-Net training using Kaapana. 138
5.31 Distribution of DICE scores for the different algorithms and datasets. . 139
5.32 Loss as a function of the training epochs for the centralized and federated trained 2D and 3D nnU-Nets for the experiments on the "seen" data. . . 142

B.1 Screenhot of the annotations of the TRC and associated certainty of the assignment. 173
B.2 Screenhots of the evaluation of the system-generated conclusions. . . . 174

List of Tables

3.1 Example translation of the different TRCs to a SOR as executed by the software for structured oncology reporting. 46
3.2 Site-specific technical parameters of the multi-site prostate MRI segmentation dataset. 47

5.1 The PCC for the validation and test set of the different approaches, which are presented in detail in Figure 4.1. 88
5.2 Number of samples for the training, validation and test set, the average number of sentences per section after sentence segmentation and the TRC counts per dataset. 93
5.3 Recall scores of the ROUGE metrics for the three key sentence extraction approaches. 94
5.4 ROUGE-1 F1 scores for the different summarization methods on the test as well as on the subsets of the four TRC. 94
5.5 Accuracies for the summarization methods and different TRC. 95
5.6 Anonymized sample SOR with the general information section and the findings section. 98
5.7 Reference conclusions and conclusions generated by the four text summarization methods. 99
5.8 Best hyperparameter settings for the four NLP models after maximizing the weighted F1 score on the n^{th} validaiton fold using $k = 5$ folds. . . . 116
5.9 Patient characteristics for the different datasets. 117
5.10 Results of the lexical complexity analysis. 118
5.11 Expected Calibration Error (ECE) loss of the four NLP models and the seven annotators on the different datasets. 118
5.12 TRC classification results for the three human annotator groups on the FTORs 119

5.13 TRC classification results of the four NLP models on the FTORsand the SOR_{Test}. . . . 120
5.14 AUC scores of the BERT model and the Linear-SVC on the three test datasets. . . . 123
5.15 DICE (%) and ASD (mm) scores for all experiments along with scores reported in works for existing methods. . . . 137

6.1 Comparison of the similarity scores as predicted by the Voting Regression (approach A1) and the medication graph (approach A3) 144

B.1 Results of the lexical complexity analysis for the comparison datasets. . 175
B.2 TRC classification results for the human annotators and the three human annotator groups on the FTORs 176
B.3 List of cases used for testing in the experiments on the seen data from the pre-processed multi-site prostate MRI segmentation dataset 177
B.4 Mean DICE (%) scores with standard deviation for each dataset, algorithm, architecture, and experiment. . . . 177
B.5 Mean ASD (mm) scores with standard deviation for each dataset, algorithm, architecture, and experiment. . . . 178

1 Introduction

1.1 The bigger picture

The outbreak of the Coronavirus disease (COVID-19) has drawn special attention to healthcare systems all over the world. The use of artificial intelligence (AI) algorithms could address major challenges occurring in such a crisis like managing limited healthcare resources, developing personalized treatment plans, or predicting virus spread (Klumpp et al., 2021). Advances in AI will more and more reshape medicine and potentially improve the experience of both clinicians and patients (Rajpurkar et al., 2022). Oncology, in particular, is at the forefront of the transformation brought by AI given the wide range of applications and the exponential growth in data (Huang et al., 2020; Kann et al., 2021). A key driving factor for AI to reach its full potential in the medical domain is its application across multiple clinical sites (Rieke et al., 2020). Only by training an AI model on sufficiently large, heterogeneous, and well-curated data, the trained model can achieve clinical-grade accuracy. Therefore, by training a model in a multi-institutional setting, it is ensured that the resulting model is safe, fair, and equitable and that it performs robustly and generalizable on unknown, unseen data (Rieke et al., 2020). The highly sensitive and tightly regulated use of health data limits data sharing and pooling, which requires working in multi-institutional clinical settings to obtain access to large, heterogeneous datasets (Rieke et al., 2020). Besides beneficial training conditions for algorithms, multi-institutional clinical settings are also essential to allow an evaluation of developed AI algorithms on a wide range of versatile, diverse clinical data (Rajpurkar et al., 2022). Independent of multi-institutional settings, Klumpp et al. (2021) divide the most promising applications of

AI in hospitals and the healthcare sector into three main areas: diagnosis, care, and logistics.

In the care area, improving healthcare efficiency and treatments through the application of AI is an important building block to handle, for example, the increasing number of older, frailer, multi-morbid patients with chronic diseases (Klumpp et al., 2021).

In the logistics area, AI can be applied, exempli gratia (e.g.), to optimize scheduling and transportation planning. As a major logistics challenge in the field of radiation therapy, Huynh et al. (2020) identify the shortages of workforce induced by the growing complexity of radiation therapy workflows. Those workflows require time-consuming, manual input by a diverse team of healthcare professionals consisting of radiation oncologists, medical physicists, medical dosimetrists, and radiation therapists. In addition, the authors fear global inequalities in cancer care between healthcare systems due to differences in available technical and infrastructural resources, as well as a gap in the knowledge and experience of the medical staff.

This work mainly focuses on potential applications of AI in the field of diagnosis. In diagnosis, automated decision support applications can help the physician in his daily clinical routine. Rising numbers of generated medical data, which include molecular assays, Electronic health record (EHR), and clinical and pathological images, pose new challenges for physicians (Klumpp et al., 2021). Especially in radiology and oncology, the amount of radiological imaging exams is growing. Huang et al. (2020) reports that an average radiologist would need to interpret an image every 3-4s over an 8-h workday to meet the increased demanded workload. Possible use cases of AI in radiology include pattern recognition, decision-making, segmentation, the extraction of biomarkers and radiomics features, or the prediction of tumor growth (Huynh et al., 2020; Kleesiek et al., 2020). Furthermore, it is almost impossible for physicians to manually examine the large amount of daily generated textual data, which contains valuable information for medical decision-making processes and further medical treatment. This holds in particular since most of the textual data is unstructured and often written in a free-form narrative language, making automatic access to the data difficult. From a medical point of view, unstructured text data also increases the risk of incompleteness and lack of comprehension of relevant information (Wang et al., 2018a; Liu et al., 2019a).

An introduction of structured reporting and the use of Natural Language Processing (NLP) can help to process a huge amount of textual data automatically (Wang et al., 2020, 2018b; Weber et al., 2020; ESR, 2018; Nobel et al., 2020). For example, the NLP task semantic textual similarity (STS) has the potential to ease clinical decision processes (e.g. by highlighting crucial text snippets in a report), to query databases for similar reports, to assess the quality of reports or for the use in question answering applications (Kades et al., 2021). Also, the downstream task of text classification has the potential to index reports and automatically retrieve information from electronic health records (Fink et al., 2022a; Yim et al., 2016). Besides the growing amount of data, also its complexity and heterogeneity, as well as the diversity in diseases, make it more difficult for physicians to include all information available for the diagnosis (Klumpp et al., 2021; Davenport and Kalakota, 2019). The increase of complex, heterogeneous medical data paired with a limited number of experts and time-consuming tasks results in a constantly growing demand for applications of AI for diagnosis (Davenport and Kalakota, 2019; Dash et al., 2019; Scherer, 2022). Watson et al. (2019) prognoses that the application of machine learning algorithms might even radically improve the ability to diagnose and treat diseases. Huynh et al. (2020) suggest that replacing time-consuming tasks like visual perception or pattern recognition could improve the availability and quality of cancer care worldwide. Finally, personalized medicine will be more easily facilitated through the use of AI (Kleesiek et al., 2020).

The following section gives an overview and background information on the technical clinical landscape as well as of AI algorithms in the clinical domain. Section 1.3 motivates the work by presenting challenges, and possible solution approaches to put AI into clinical practice. The main contributions of this work are introduced in detail in Section 1.4.

1.2 Status quo

1.2.1 The data landscape in the clinics

The purpose of healthcare systems is to prevent, diagnose and treat health-related issues or impairments of human beings (Dash et al., 2019). Generally, a healthcare system consists of health professionals like physicians or nurses and health facilities like clinics and hospitals (Dash et al., 2019). Throughout the treatment of patients, a lot of heterogeneous data is generated, partially by health professionals, e.g., in the form of narrative reports or by devices and equipment used to assess a patient's health state. The first medical report dates back to 1600 BC in Egypt, written on papyrus text (Dash et al., 2019; Gillum, 2013). The format of handwritten or typed reports persisted until the advent of computer systems. A standard digital format to store medical data called EHR was introduced in 2003 by the division of the National Academies of Sciences, Engineering, and Medicine to improve the healthcare sector for the benefit of patients and clinicians (Dash et al., 2019). As of 2012, it is reported that 500 petabytes of EHR data were generated and Sun and Reddy (2013) prognosed to reach 25 exabytes of data by 2020 (Yu et al., 2019).

EHRs contains past, present, and future information about the patient's health condition. The longitudinal information includes, amongst others, patient demographics, vital signs, medications, laboratory data, imaging data, and narrative text data (Shamout et al., 2021). To store the various data, healthcare institutions have a variety of healthcare information systems at their disposal which include: Hospital Information System (HIS), Clinical Information System (CIS), Picture Archiving and Communication System (PACS), Laboratory Information System (LIS), and Radiology Information System (RIS), with each one saving specific types of clinical data (Yu et al., 2019). In general, clinical data can be divided into medical images, clinical notes, and other data. Other data include e.g., results of physiological measurements like lab results or vital signs, demographic information or payment and insurance information (Yu et al., 2019). In the following, medical image data and clinical notes are discussed in more detail.

Medical images are recorded by different techniques, which include, e.g., X-rays, computed tomography (CT), magnetic resonance imaging (MRI), Optical Coherence Tomography (OCT), microscopy imaging, and positron emission tomography (PET) (Yu et al., 2019). The type of imaging denotes the modality of the recorded image. Each technique has its advantages and disadvantages. According to Yu et al. (2019), MRIs are suited to detect pathologies of the brain, of the cardiac system, and of the bones and joints. CTs are good when examining abdominal organs and the chest. X-rays are often used for the chest and the breast. In 1985, the image format Digital Imaging and Communications in Medicine (DICOM) was introduced as a standard image format in radiology and nuclear medicine by the National Electrical Manufacturers Association (NEMA) (NEMA, 2021). For the application of deep learning (DL) or machine learning models, other formats are often preferred by the developers, which include Neuroimaging Informatics Technology Initiative (NIfTI) (Initiative, 2011) or "nearly raw raster data" (Nrrd) (teem, 2023) (Scherer, 2022). Medical images are typically 2D or 3D recordings of the body. In the case of 3D, multiple 2D slices of the body are recorded and put together into one image. There exist also 4D medical images, which include a temporal axis in the image or different modalities. The images are saved in vector format, where 2D images consist of pixels and 3D images of voxels (Yu et al., 2019). To handle segmentation objects in a standardized way, e.g., the segmentation of a liver, the DICOM for Quantitative Imaging (dcmqi) tool from Andrey Fedorov (2023) offers ways to create standard-compliant DICOM Segmentation Objects (DICOM SEG) from a NIfTI or Nrrd mask (Scherer, 2022). In hospitals and clinics, all imaging data generated by the different medical devices are stored in a dedicated PACS. Along with the image information, metadata about the patient as well as technical parameters are captured. There exist many tools to examine and analyze these images or to create pixel-wise annotations, examples include the Medical Imaging Interaction Toolkit (MITK) (Wolf et al., 2004; Nolden et al., 2013), 3D Slicer (Lasso, 2019) or zero-footprint web-based Open Health Imaging Foundation (OHIF) Medical Imaging Viewer (Urban et al., 2017).

Clinical notes are often written in narrative text form and contain discharge summaries and various kinds of measurement reports (Yu et al., 2019). Discharge summaries can include a description of lab test results, physician diagnoses, drugs, and treatments. Indirect information like family history, medical history or allergies can also be part of the report (Yu et al., 2019). In many cases, the reports are composed of acronyms and nonstandard clinical jargon and do not follow institution-specific document structures (Hasan and Farri, 2019). Furthermore, clinical reports often differ in quality and length due to time limitations or because snippets are simply copy-pasted from other reports (Kades et al., 2021; Embi et al., 2013; Zhang et al., 2014).

Multiple efforts exist to standardize narrative reports to improve readability and automatic information extraction. For example, in the field of radiology, Weber et al. (2020) introduced Structured Oncology Reporting at University Hospital Heidelberg (UKHD) to increase the completeness and comprehensibility of reports. The concept is presented in Section 3.2.2 in more detail. Besides the standardization of the data itself, efforts exist to standardize the communication and transfer of data by introducing well-defined application programming interfaces (APIs) between different clinical software applications. A promising future solution might be a SMART on FHIR protocol using the HL7 standard (Mandel et al., 2016; SMART, 2019; HL7, 2019; Cutillo et al., 2020; Scherer, 2022).

1.2.2 The algorithmic landscape in the clinics

The concept of artificial intelligence (AI) exists since the 1950s. The original definition states that a machine can perform a given task just as well as a human (Kann et al., 2021). Over the decades the field has transformed from rule-based systems to Neural networks (NNs) and deep neural networks (DNNs), mainly driven by the exponential growth of data and the computing hardware advances (Huynh et al., 2020; Kann et al., 2021). Also in medicine, with the explosion of healthcare data, especially, DL has gained popularity for its ability in feature representation and pattern recognition (Yu et al., 2019).

Rule-based systems, machine learning, neural networks, and deep neural networks

Very early AI systems were rule-based systems, id est (i.e.), the process of such a system is solely defined by heuristic rules given by human experts. The wide use of rule-based systems started in the 1980s, especially for tasks like clinical decision support (Davenport and Kalakota, 2019). While the systems perform well on a given task and are easy to understand, they become more complicated with an increasing number of rules and edge cases and often lack generalizability (Huynh et al., 2020; Davenport and Kalakota, 2019). Over the decades, rule-based systems have been more and more replaced by machine learning (ML) algorithms.

ML algorithms are based on statistical methods to fit models to data. They are applied to supervised or unsupervised tasks. In supervised tasks, a model learns a mapping between an input and an output based on labeled training data. In unsupervised tasks, the model is trained to identify patterns and properties of an underlying distribution in the data automatically (Shickel et al., 2018; Davenport and Kalakota, 2019). Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Conditional Random Fields (CRF), Bayesian Networks (BN), Principal Component Analysis (PCA), and Latent Dirichlet Allocation (LDA) are statistical machine learning methods, which find manifold applications when working with clinical data (Yu et al., 2019).

The use of NNs started in the 1960s with the mathematical development of the backpropagation algorithm (Davenport and Kalakota, 2019; Huynh et al., 2020), which is responsible for updating the weights of a NN based on a given loss function (Huynh et al., 2020). With faster Graphics Processing Units (GPUs) and increasing computing power, DNNs have evolved. DNNs are characterized by multiple intermediate hidden layers between input and output layers and are able to learn very complex, non-linear relationships in data (Huynh et al., 2020). DL is widely used in radiology and oncology. Popular tasks include classification, object detection, semantic segmentation, image processing, and NLP (Davenport and Kalakota, 2019; Ueda et al., 2019). Since this work mainly covers NLP and medical imaging algorithms, recent developments in these algorithms are discussed in more detail below.

Natural Language Processing algorithms

For the field of healthcare, the term clinical NLP evolved. The use of clinical NLP involves all kinds of applications in conjunction with clinical notes. One major application is automated information extraction from clinical notes. The extracted information can be used to index clinical notes to query the data automatically or to transfer key information into a tabular form (Velupillai et al., 10.03.2018). Further topics of interest are text generation, summarization and classification as well as semantic textual similarity, question answering, named entity recognition and de-identification (Davenport and Kalakota, 2019; Shickel et al., 2018).

The field of NLP followed a similar transition from rule-based systems over statistical machine learning methods to NNs and DNNs. Traditional rule-based systems in clinical NLP work based on heuristic rules defined by a medical expert. For example, in Kang et al. (2012), they introduce a NLP module consisting of five submodules, with each one defining rules to improve biomedical concept normalization, which describes the mapping between a clinical text and a biomedical knowledge base. Rules can be hard-coded or enhanced by methods such as string matching and the application of Regular expression (regex) or more sophisticated machine learning methods (Kang et al., 2012).

A major challenge when working with texts is their representation as vectors. Common ways to create input representations are Bag-of-Words (BOW), one-hot vector encoding techniques (Hasan and Farri, 2019), or term frequency-inverse document frequency (TF-IDF) scorings. Word2Vec (Mikolov et al., 2013) and Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) are early-developed methods that make use of NNs. These methods capture the context of the words by creating a word representation based on its surrounding. Training strategies for Word2Vec include, e.g., Skip Gram and Continuous BOW (Kang et al., 2012). The models are commonly pre-trained on many data using self-supervised tasks, such as mask word prediction. Then, they are further fine-tuned using so-called downstream tasks such as semantic textual similarity or text classification. DL algorithms learn high-dimensional vector-representation that are based on character-level n-grams, words, phrases, sentences or documents (Kang et al., 2012). Common DL network architectures used in NLP are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Long short-term memorys (LSTMs), where the latter two are widely used for the task of machine translation.

As summarized in Kades et al. (2021), the field of NLP was revolutionized by the invention of transformer models in recent years. With the use of the attention mechanism (Vaswani et al., 2017), the transformer architecture makes it possible to leverage vast amounts of unlabeled data by capturing its semantic knowledge and generating a model that contains a universal language representation. A well-known model introduced in 2018 is Bidirectional Encoder Representation from Transformers (BERT), which is pre-trained on a massive amount of data with two unsupervised tasks: (1) next sentence prediction and (2) masked word prediction (Devlin et al., 2019). After pre-training, the model can be extended by additional layers and further fine-tuned on respective downstream tasks like STS, text classification, or text summarization. Depending on the task and the purpose of the transformer-based model, many new architectures have evolved. Those either incorporate varying pre-training tasks (Clark et al., 2020; Devlin et al., 2019; Sun et al., 2019), make use of multitask learning approaches (Liu et al., 2019b) or combine the two approaches (Raffel et al., 2020; Kades et al., 2021). To adapt to domain-specific datasets like clinical notes, many efforts are made, such as further pre-training existing models or training models from scratch solely on data of the target domain (Alsentzer et al., 2019; Lee et al., 2020; Shrestha, 2021).

NLP was successfully applied in the clinical domain, e.g., for the automatic extraction of International Classification of Diseases (ICD)-codes from clinical notes (Lita et al., 2008; Koopman et al., 2015) or for automatically extracting clinically relevant oncologic information from radiologic reports (Kehl et al., 2019).

Medical image computing

Medical image computing (MIC) refers to the application of algorithms on medical images. According to Ueda et al. (2019) and Scherer (2022), the most widespread tasks include classification, object detection, semantic segmentation, image processing, and image registration. Also in MIC early methods include rule-based systems, pixel-wise operations, and the usage of statistical ML methods. The DL methods used in MIC are mostly built on top of methods from the field of Computer Vision (CV). The most common model architecture in MIC is the CNN. Recently, visual transformer architectures have found their way into MIC. A well-known network for the task of semantic segmentation is the U-Net architecture proposed by Ronneberger et al. (2015). Due to its architecture, which consists of up- and down-sampling layers, it can capture the full context of an image (Yu et al., 2019). In conjunction with the Medical Segmentation Decathlon (Medical Segmentation Decathlon, 2023) in 2018, Isensee et al. (2021) introduced the no new net U-Net (nnU-Net), which uses the U-Net as proposed by Ronneberger et al. (2015), but optimizes the pre-processing of the images, the architecture of the network, the training and the post-processing based on characteristics of the underlying dataset. More information on the nnU-Net is given in Section 3.1.2.

Applications of MIC that can be used directly in clinical practice are, for example, the prediction of a 3-year lung cancer risk from CTs, worked out by Huang et al. (2019), or an implementation of an end-to-end CNN architecture on diffusion-weighted magnetic resonance (MR) images, proposed by Jäger et al. (2017). The model is able to predict whether an invasive biopsy for a breast cancer patient is really necessary.

1.2.3 Real-world initiatives between multiple clinics

There exist many initiatives which operate in multi-institutional settings. This section will introduce two German initiatives, namely the Joint Imaging Platform (JIP) (Joint Imaging Platform, 2023) and Radiological Cooperative Network (RACOON) (The radiology cooperation in NUM, 2023), which are of particular interest in this work because most of the presented problems are motivated and implemented within the two initiatives. They aim to put potential applications of AI in healthcare, presented in Section 1.1, into practice as well as to tackle respective challenges, presented in Section 1.3. Some of the questions and problems presented in this work are motivated by requirements and use cases within those initiatives. Details about the technical setup within the initiatives are given in Section 3.1.2. The following two paragraphs are strongly based on the introduction of the projects from the work of Scherer (2022).
The Joint Imaging Platform (JIP) (Joint Imaging Platform, 2023) is a strategic initiative to establish a standardized, distributed information technology (IT) infrastructure for medical imaging in cancer research across multiple clinical centers in Germany. The initiative was launched in 2017 by the German Cancer Consortium (DKTK) and includes 11 cooperation partners whose locations are shown in Figure 1.1: Charité Berlin, Dresden, Düsseldorf, Essen, Frankfurt, Freiburg, Heidelberg, Mainz, LMU Munich, TU Munich, and Tübingen. Besides the technical infrastructure, the project also aims at improving collaborations and networking of experts as well as the creation of joint projects in MIC. Since 2017, a technical infrastructure, which is also called JIP, was implemented and distributed to each site. The JIP embeds itself into the clinical IT infrastructure without disturbing the clinical routine. It has an access point to receive DICOMs from the clinical PACS, offers a meta-data-based cohort selection and most importantly allows the execution of state-of-the-art algorithms. Following Scherer (2022), the platform should give support for the execution of clinical studies by allowing for image-based patient stratification, therapy monitoring, radiomics analysis as well as early detection and progression assessment. Currently implemented workflows include, e.g., the nnU-Net and a radiomics pipeline. Unique selling points of the JIP comprise from a technical point of view, its seamless integration into the clinical IT infrastructure, its highly standardized execution of algorithms using container technologies as well as the possibility to evaluate and train AI algorithms across different clinical centers. From a political point of view, its particular value is the creation of a collaboration

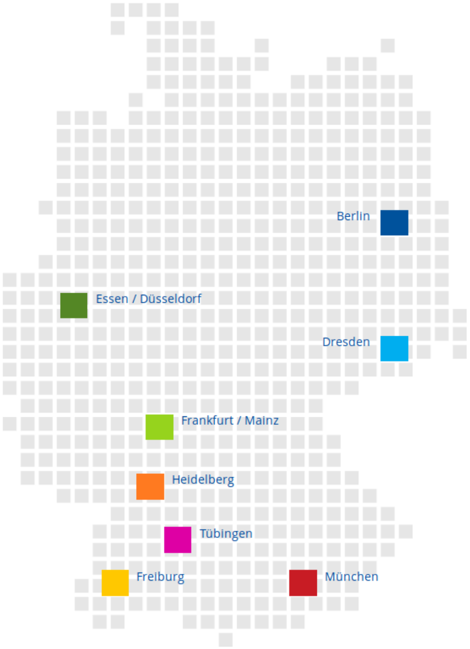


Figure 1.1: *Locations of the cooperation partners within the Joint Imaging Platform (JIP) initiative. This figure was adapted from Scherer (2022).*

between equal partners with each one maintaining sovereignty over their data and at the same time sharing expertise with other partners (Scherer, 2022; Scherer et al., 2020).

Radiological Cooperative Network (RACOON) (The radiology cooperation in NUM, 2023) was launched in 2020 as a response to the COVID-19 pandemic from the Network of University Medicine (NUM). The network comprises all 36 university hospitals in Germany. The locations of the sites are illustrated in Figure 1.2. The main driving factor for NUM to initiate the RACOON project is to handle pandemic situations in a better, clinical over-arching way. RACOON was funded in the first year with 150 million euros from the German Federal Ministry of Education and Research (BMBF). Besides thirteen other funded projects in response to the COVID-19 pandemic, RACOON has the specific target to tackle the radiological aspects of the pandemic. Radiological imaging is well suited to monitor and assess pulmonary diseases, which occur in conjunction with severe SARS-CoV-2 infections. Similar to the JIP, RACOON offers the unique possibilities to connect multiple clinical centers, to establish standards in reporting, image annotations and the execution of AI algorithms as well as to create a clinic over-arching collection of well-annotated and reported radiological data. By connecting the clinics with a central node, the setup will allow training and

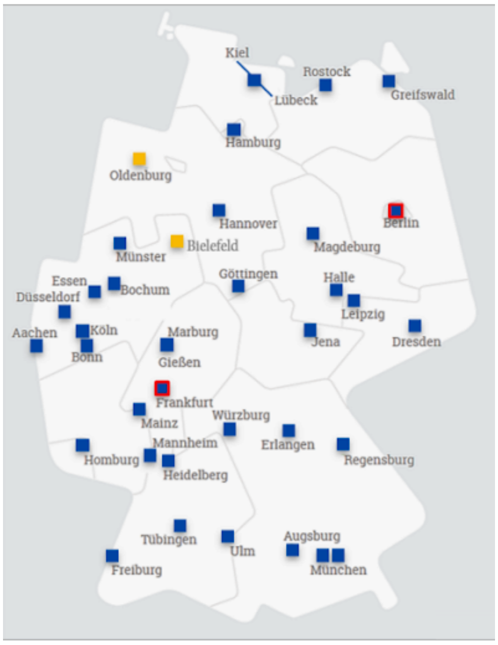


Figure 1.2: *Locations of the cooperation partners within the Radiological Cooperative Network (RACOON). This figure was adapted from Scherer (2022).*

evaluation of AI algorithms on a diversity of data from different clinical sites and patient cohorts. The university hospitals Charité Berlin and Frankfurt are leading the project in close cooperation with the German Cancer Research Center (DKFZ), the Technical University of Darmstadt (TUDa), the Fraunhofer Institute for Digital Medicine (MEVIS), the ImFusion GmbH and the Mint MedicalGmbH, who are amongst others responsible for the legal aspects and the deployment of the systems.

In the first phase of RACOON, structured reports were collected with the help of the tool mintLesion from the Mint Medical GmbH for a cohort of COVID-19 patients at each clinical site. Additionally, COVID-19 relevant anatomies and pathologies in the lung were segmented with the help of the segmentation framework SATORI from MEVIS and ImFusion Labels from ImFusion. Finally, segmentation algorithms were trained with the help of a JIP similar infrastructure provided by the DKFZ. As mentioned above, more details on the technical setup are given in Section 3.1.2.

1.3 Open challenges for the application of algorithms in multi-institutional clinical settings

Kleesiek et al. (2020) conclude that despite the existence of many AI algorithms in oncology imaging, their widespread application in clinical practice is still very limited. The following section gives a broad overview of problems and challenges that are responsible for the gap between algorithms and their application in clinical practice. The next section suggests possible solutions, approaches and first steps toward resolving those challenges. The challenges and potential solution approaches are illustrated in Figure 1.3. The following content is a result of literature research including the works from Cutillo et al. (2020), Davenport and Kalakota (2019), Hasan and Farri (2019), Huang et al. (2020), Huynh et al. (2020), Kaissis et al. (2020), Kann et al. (2021), Kleesiek et al. (2020), Klumpp et al. (2021), Rajpurkar et al. (2022), Reyes et al. (2020), Rieke et al. (2020), Shickel et al. (2018) and Watson et al. (2019).

1.3.1 Challenges

According to Rajpurkar et al. (2022), three main fields hinder the application of AI algorithms in healthcare and clinics: Accountability and regulations, fairness as well as implementation challenges.

In the accountability and regulations area, multiple unsolved questions and critics arise. Rajpurkar et al. (2022) point out that AI systems might be in real-world settings less helpful than suggested or that the systems may be too slow or complicated for the end-user. In addition, up to now AI technologies are most of the time classified as "software as a medical device" by national and international regulatory bodies, which means that certain regulatory standards need to be fulfilled (Huynh et al., 2020). However, so far there exist no common regulations and ways of how to validate that AI algorithms are robust and generalizable across different clinical centers and patient populations or that the algorithms do not impact treatment in a negative way (Rajpurkar et al., 2022). To facilitate such clinical evaluations, AI algorithms might need to be integrated into an existing system, and the AI systems would need to follow certain standards such that data privacy and security are guaranteed. Since the usefulness of AI systems can be heavily influenced by the end-user, adequate, regulated training for healthcare professionals might be necessary (Klumpp et al., 2021; Rajpurkar et al., 2022).

From a fairness point of view, missing guarantees for ethical usage of data from actors and stakeholders pose challenges for a wide usage of AI algorithms in healthcare and clinics, especially, because high-quality data and algorithms have a significant business value (Rieke et al., 2020; Rajpurkar et al., 2022). Other issues relate to the protection of patient privacy (Rajpurkar et al., 2022). For example, it is possible to reconstruct a patient's face just from a CT or MRI image (Rieke et al., 2020). Finally, the data and as a consequence also the trained models might be biased towards certain population groups, which would result in unequal treatments of patients (Rajpurkar et al., 2022; Klumpp et al., 2021).

Therefore, implementation challenges also include a potential bias in the dataset and missing standardizations in data across clinical centers (Rajpurkar et al., 2022; Klumpp et al., 2021; Weber et al., 2020; Willemink et al., 2020). The quality of a dataset is often lowered by missing standards for the data acquisition, which then often requires laborious curation before it can be used for the development of AI systems (Huynh et al., 2020). As discussed in Section 1.2.2, especially narrative reports pose unique challenges for automated processing due to their complex, individual and inconsistent document structure and organization (Hasan and Farri, 2019). Additionally, the development of robust and well-performing AI algorithms for a broad application in clinics is hindered by a lack of heterogeneous high-quality annotated benchmark datasets, mostly because the annotation of datasets requires experts and is very time-consuming (Rajpurkar et al., 2022; Huynh et al., 2020; Shickel et al., 2018; Willemink et al., 2020). Furthermore, many research groups and industries have only access to data of limited sample size and from small geographic areas (Willemink et al., 2020). Finally, the training and validation of an AI model on data across multiple institutions, which might be the final step for the transition from research to clinical practice, is often hindered by data silos, privacy concerns as well as legal and ethical requirements to protect the patient's privacy (Rieke et al., 2020; Kaissis et al., 2020; Willemink et al., 2020).

1.3.2 Solution approaches

Based on the presented challenges, many potential solution approaches arise for tackling the presented challenges. They can be categorized into four fields: Standardization and curation, clinical adaptation, validation and trust, as well as explainability and interpretability.

Standardization and curation

Multiple attempts already exist to standardize data as well as to provide more data for public usage. This includes the introduction of the DICOM standard along with the efforts from Andrey Fedorov (2023) to create standard-compliant DICOM Segmentation Objects. Another example is the Cancer Imaging Archive (TCIA) (Clark et al., 2013), which promotes data sharing (Huynh et al., 2020). Also, the introduction of EHR was a huge step toward standardizing clinical data. However, in practice, EHRs often lack interoperability, which is why the development of more standardized clinical APIs such as Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) is important (HL7, 2019; Cutillo et al., 2020). In addition, important solution approaches include tackling data-related issues such as missing data, secondary data sources integration, missing data structure, and more informative metadata (Cutillo et al., 2020). As emphasized by Kann et al. (2021), the introduction and compliance of guidelines such as the Findable, Accessible, Interoperable and Reusable (FAIR) (FAIR Principles, 2023) principles are important next steps to enhance the application of AI algorithms as well as the reusability of data (Wilkinson et al., 2016). As discussed in Section 1.2.2, the introduction of Structured Oncology Reports (SORs) by Weber et al. (2020) over narrative reports are examples of such initiatives. Another important step to increase the usability of data in AI applications is the development and usage of curation tools, which can help data cleaning steps and cohort selections (Huynh et al., 2020; Kades et al., 2022a; Willemink et al., 2020). Also, for enabling the development of robust and well-performing algorithms, the generation of sufficiently large, curated, high-quality annotated datasets is of high importance (Rajpurkar et al., 2022; Huynh et al., 2020; Willemink et al., 2020). Besides the standardization and provision of well-curated data, standardizations in developing, executing, and deploying algorithms in clinical settings are important future steps (Willemink et al., 2020). Likewise, for federated scenarios and data sharing across hospitals, standardized infrastructures are essential (Rieke et al., 2020). Software projects like MONAI (Medical Open Network for Artificial Intelligence) and dcmqi (DICOM for Quantitative Imaging) as well as infrastructure projects like NVIDIA Clara (NVIDIA Clara, 2023) and Kaapana (Scherer et al., 2023) are examples of such efforts. NVIDIA Clara is a platform to build manage and deploy intelligent medical imaging workflows and instruments and Kaapana is an open-source toolkit for platform provisioning in the field of medical data analysis.

Clinical adaptation

Considering an increasing application of AI algorithms in clinics, multiple authors suggest that it will be important to train and educate healthcare professionals to be able to effectively use, interpret and understand the limits of AI applications in their daily clinical practice (Huynh et al., 2020; Davenport and Kalakota, 2019; Klumpp et al., 2021). Furthermore, it will be important that humans and AI algorithms work together and that healthcare professionals fully understand the benefits of AI applications (Rajpurkar et al., 2022; Klumpp et al., 2021). This concludes that the skillset of healthcare professionals needs to be extended towards data science. In addition, the usability of AI algorithms is of great importance for their acceptance in clinical practice. Therefore, the development of AI systems requires strong collaborations between patients, physicians, computer scientists, data scientists, and policymakers (Cutillo et al., 2020; Kleesiek et al., 2020). In addition, the algorithms themselves need adaptations for clinical usage. Therefore, it needs to be investigated, whether state-of-the-art algorithms developed in non-medical domains also perform best in clinical settings. Chen et al. (2019) argue, for example, that conventional ML methods often provide simpler, computationally cheaper, and more useful methods than DL for data modeling tasks. Finally, Huang et al. (2020) emphasize the need to combine data from multiple sources for a successful application of AI since considering all generated data is necessary to understand the full clinical context and to come to a correct clinical decision. Besides laboratory data, combining the information from images and their corresponding narrative reports is essential. Huang et al. (2020) present e.g., early, joint, and late fusion techniques for incorporating multiple data sources in an AI system.

Validation of artificial intelligence methods

Elaborating the need for clinical adaptation also includes a widespread validation and assessment of AI algorithms on real, heterogeneous, and more multi-institutional datasets (Rajpurkar et al., 2022). Often medical AI algorithms are developed and optimized within organized challenges or on small in-house datasets. However, this does not automatically guarantee the same performance of the model on real-world data, e.g., due to learned biases (Peiffer-Smadja et al., 2020). For example, It might be challenging to predict the performance on edge cases not part of the training data (Kann et al., 2021). Increasing the amount of publically available clinic data such as the UK Biobank (UK biobank, 2023) and TCIA (Clark et al., 2013), as well as more clinical benchmark datasets such as the National NLP Clinical Challenges (N2C2) NLP Research datasets (DBMI Data Portal, 2023), the Medical Segmentation Decathlon (Medical Segmentation Decathlon, 2023) or the Brain Tumor Segmentation (BraTS) dataset are promising directions to make methods more robust (Shickel et al., 2018; Peiffer-Smadja et al., 2020). Politically, another direction is to establish more national and international collaborations similar to RACOON and JIP to enable data access and, therefore, the execution of algorithms on real-world data. These directions should be paired with the design of more multi-institutional clinical studies, which use AI algorithms. This is important because, while much research focuses on improving the performance of algorithms, only a few studies exist to assess and validate their clinical significance and generalizability in real-world use cases and studies with real-world data (Kann et al., 2021; Rajpurkar et al., 2022; Klumpp et al., 2021; Linna and Kahn, 2022). Of course, to support this also from a technical point of view, toolkits like Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis (Kaapana) or NVIDIA Clara need to provide the necessary functionalities and have to meet the legal requirements to find their way into the clinics. Furthermore, the application and evaluation of AI systems in federated scenarios are of utmost importance to increase their robustness and generalizability.

This again requires establishing technical and political solutions for federated learning scenarios and validations (Rieke et al., 2020; Kaissis et al., 2020).

Trustworthiness, explainability and interpretability

The performance gain of DL models often comes at the cost of limited explainability and interpretability (Zhu et al., 2016; Kann et al., 2021). Especially in healthcare, it is essential to understand the reasoning behind the results of DL algorithms. Trustworthiness, explainability, and interpretability are vital for the physician to determine whether the model works as expected and for the patient to establish trust towards the use of AI algorithms (Huynh et al., 2020; Cutillo et al., 2020). Challenges that must be tackled by interpretability and explainability are, e.g., whether a model contains errors or systematic biases (Huynh et al., 2020; Cutillo et al., 2020; Peiffer-Smadja et al., 2020). However, also regulatory, ethical, and fairness aspects require the understanding of the so-called "black box" of AI algorithms to allow its usage in clinical practice (Reyes et al., 2020).

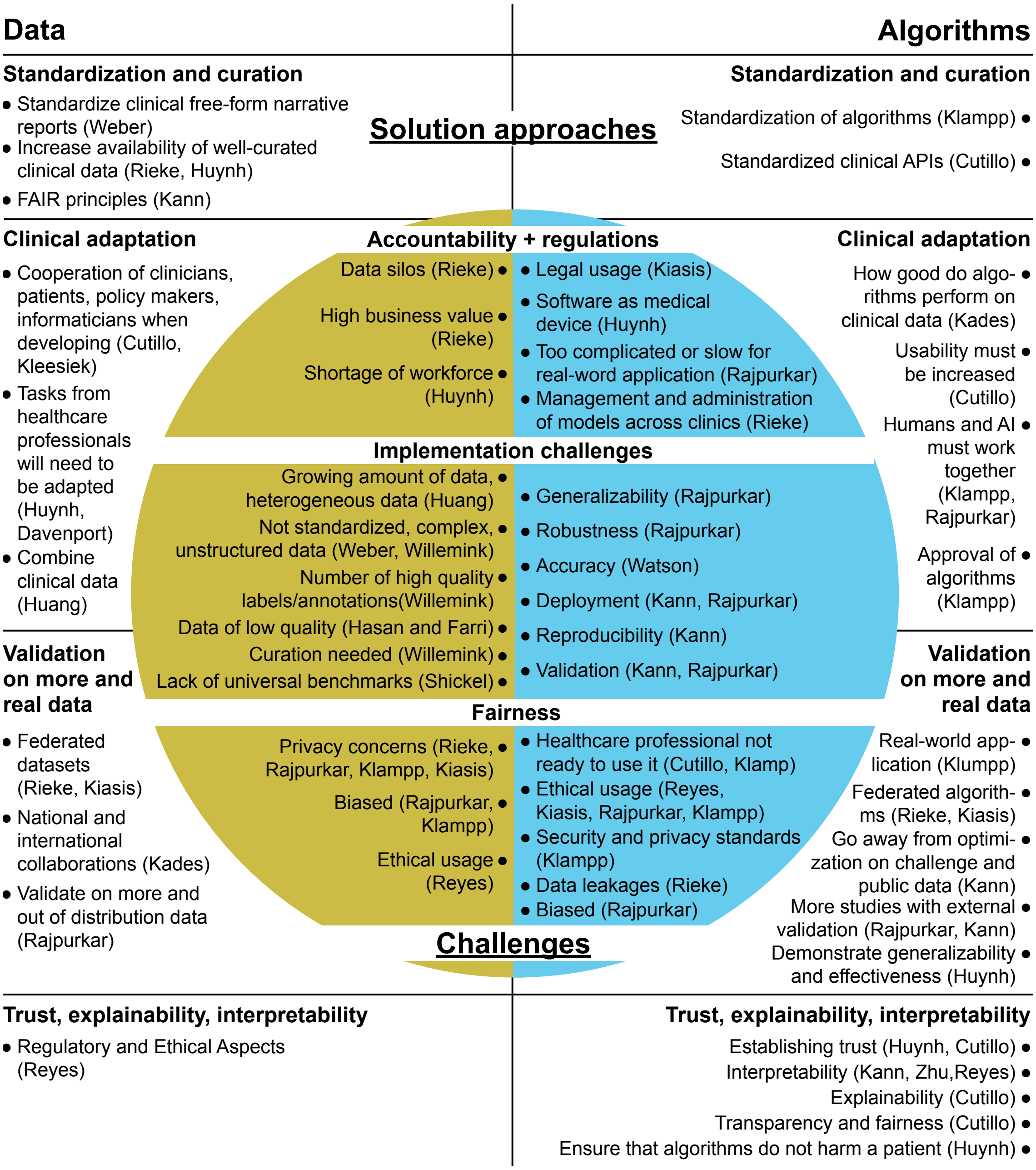

Data
Algorithms
Solution approaches
Standardization and curation
• Standardize clinical free-form narrative reports (Weber)
• Increase availability of well-curated clinical data (Rieke, Huynh)
• FAIR principles (Kann)
Standardization and curation
Standardization of algorithms (Klampp) •
Standardized clinical APIs (Cutillo) •
Clinical adaptation
• Cooperation of clinicians, patients, policy makers, informaticians when developing (Cutillo, Kleesiek)
• Tasks from healthcare professionals will need to be adapted (Huynh, Davenport)
• Combine clinical data (Huang)
Accountability + regulations
Data silos (Rieke) •
High business value • (Rieke)
Shortage of workforce • (Huynh)
• Legal usage (Kiasis)
• Software as medical device (Huynh)
• Too complicated or slow for real-word application (Rajpurkar)
• Management and administration of models across clinics (Rieke)
Clinical adaptation
How good do algo- • rithms perform on clinical data (Kades)
Usability must • be increased (Cutillo)
Humans and AI • must work together (Klampp, Rajpurkar)
Approval of • algorithms (Klampp)
Implementation challenges
Growing amount of data, • heterogeneous data (Huang)
Not standardized, complex, • unstructured data (Weber, Willemink)
Number of high quality • labels/annotations(Willemink)
Data of low quality (Hasan and Farri) •
Curation needed (Willemink) •
Lack of universal benchmarks (Shickel) •
• Generalizability (Rajpurkar)
• Robustness (Rajpurkar)
• Accuracy (Watson)
• Deployment (Kann, Rajpurkar)
• Reproducibility (Kann)
• Validation (Kann, Rajpurkar)
Validation on more and real data
• Federated datasets (Rieke, Kiasis)
• National and international collaborations (Kades)
• Validate on more and out of distribution data (Rajpurkar)
Fairness
Privacy concerns (Rieke, • Rajpurkar, Klampp, Kiasis)
Biased (Rajpurkar, • Klampp)
Ethical usage • (Reyes)
• Healthcare professional not ready to use it (Cutillo, Klamp)
• Ethical usage (Reyes, Kiasis, Rajpurkar, Klampp)
• Security and privacy standards (Klampp)
• Data leakages (Rieke)
• Biased (Rajpurkar)
Challenges
Validation on more and real data
Real-world app- • lication (Klumpp)
Federated algorith- • ms (Rieke, Kiasis)
Go away from optimi- • zation on challenge and public data (Kann)
More studies with external • validation (Rajpurkar, Kann)
Demonstrate generalizability • and effectiveness (Huynh)
Trust, explainability, interpretability
• Regulatory and Ethical Aspects (Reyes)
Trust, explainability, interpretability
Establishing trust (Huynh, Cutillo) •
Interpretability (Kann, Zhu,Reyes) •
Explainability (Cutillo) •
Transparency and fairness (Cutillo) •
Ensure that algorithms do not harm a patient (Huynh) •

Figure 1.3: *Challenges and potential solution approaches for the application of algorithms in multi-institutional clinical settings. Challenges are summarized within the circle. Possible solution approaches to address those problems are collected outside of the circle. The left side concentrates on data- and the right side on algorithm-related challenges and solution approaches.*

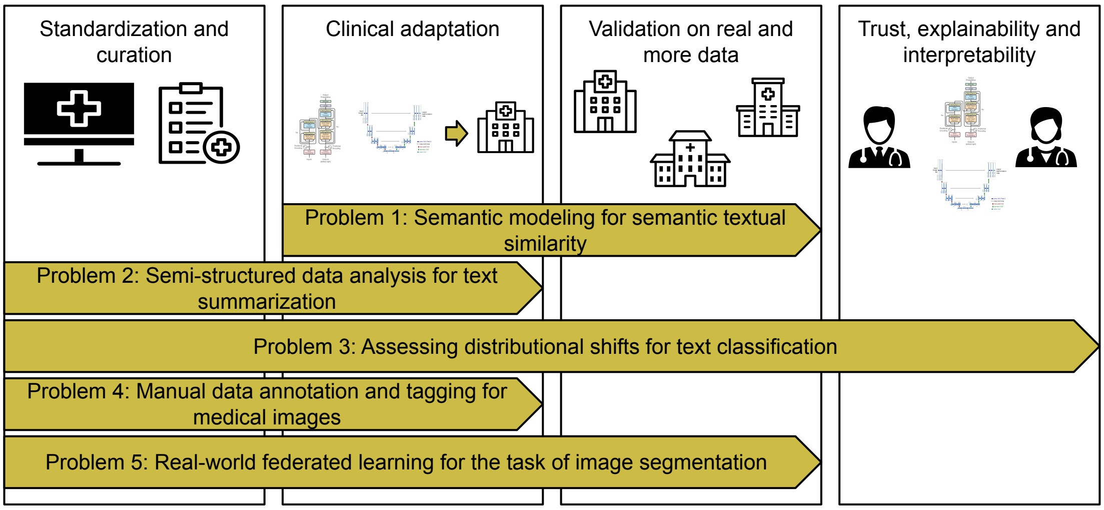

Standardization and curation
Clinical adaptation
Validation on real and more data
Trust, explainability and interpretability
Problem 1: Semantic modeling for semantic textual similarity
Problem 2: Semi-structured data analysis for text summarization
Problem 3: Assessing distributional shifts for text classification
Problem 4: Manual data annotation and tagging for medical images
Problem 5: Real-world federated learning for the task of image segmentation

Figure 1.4: *Overview of the five main contributions of this work that address challenges presented in Section 1.3 and that provide potential solutions for the application of algorithms in multi-institutional clinical settings. Figure of transformer model adapted from Vaswani et al. (2017), and figure of U-Net model adapted from Ronneberger et al. (2015). Icons by Yosua Bungaran and TkB, from thenounproject.com CC BY 3.0.*

1.4 Contributions and outline

This work discusses five selected problems of the above-mentioned challenges and potential solution approaches. Each one contributes to enabling the application of AI algorithms in multi-institutional clinical settings. Figure 1.4 gives an overview of the selected problems. The problems are introduced below and discussed in more detail in the following chapters of this work. After an overview of related work in Chapter 2, the software toolkit Kaapana and multi-institutional datasets used in this work are introduced in Chapter 3. Utilized and developed methods are presented in Chapter 4. Details on the experiments and results are given in Chapter 5, followed by a discussion in Chapter 6. The results are summarized in Chapter 7. In addition, Chapter A lists the major publications on which the work is based, along with information about the authors' contributions.

1.4.1 Semantic modeling for semantic textual similarity

The first problem tackles the task of semantic textual similarity (STS) in the clinical domain. Semantic textual similarity (STS) refers to a subtask in NLP that determines the degree of semantic similarity between two sentences or text snippets. It is used for tasks like question answering, semantic information retrieval, and text summarization (Kades et al., 2021; Cer et al., 2017; Fan et al., 2019; Majumder et al., 2016; Zhang et al., 2015; Gomaa and Fahmy, 2013). Since 2016, challenges such as the National NLP Clinical Challenges (N2C2), formerly known as i2b2 (Informatics for Integrating Biology & the Bedside) NLP Shared tasks, with different NLP downstream tasks are organized to improve the development of NLP systems on clinical and biomedical text data. STS in the clinical domain has the potential to assess the quality of clinical notes by spotting copy-pasted contents, which can increase the quality of the reports and eventually also improve the performance on downstream tasks such as information extraction. Additionally, STS can be used in decision support systems and medical question-answering applications (Kades et al., 2021). The Sections 2.1, 4.1, 5.1, and 6.1 address problems of semantic modeling for STS in the clinical domain and build heavily on the publication "Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm, Development and Validation Study (Kades et al., 2021)". The work was published as a contribution to track 1, N2C2/Open Health Natural Language Processing (OHNLP) Track on Clinical Semantic Textual Similarity, of the 2019 National NLP Clinical Challenges (N2C2). The main contributions consist of improving the performance of BERT in predicting the similarity of English clinical sentence pairs by:

- a "modification of the BERT architecture by adding additional similarity features and employing a built-in ensembling method" (Kades et al., 2021),
- introducing "a graph-based similarity approach for a subset of structured sentences in which the knowledge of the training set is extrapolated to unseen sentence pairs of the test set" (Kades et al., 2021).

Furthermore, an in-depth statistical analysis of the training and test dataset reveals statistical differences, which help to interpret the results of the different approaches.

1.4.2 Semi-structured data analysis for text summarization

This problem focuses on evaluating and improving the performance of text summarization on Structured Oncology Reports (SORs). As presented earlier, Weber et al. (2020) introduced SORs to increase the comprehensibility and completeness of imaging findings, which contain information on the diagnosis and treatment guidance for the patient's disease progression. In particular, the conclusion section of SORs, containing the assessments of the image findings, is crucial for treatment decisions and strategies. Therefore, this contribution addresses the question of whether it is possible to extract relevant information from the general information and findings section of SORs to compose the conclusion section automatically. Automatic text summarization of findings can give the physician time for better patient care. Also, it could be used, e.g., to get a summary of a patient’s history, thus helping in clinical decisions. Of course, the question arises whether all information necessary to create the conclusion is part of the general information and findings section and, if this is not the case, whether the model can extrapolate additional knowledge through training. The Sections 2.2, 4.2, 5.2, and 6.2 are based on the publication "Fine-tuning BERT Models for Summarizing German Radiology Findings (Liang et al., 2022)" as well as the master thesis of Siting Liang with the title "Summarizing German Radiology Findings for Cancer Patients (Liang, 2021)". Their main contributions include the following:

- a baseline performance of the pre-trained German BERT model to generate conclusions based on the general information and finding sections of German radiology reports (Liang et al., 2022),
- an improvement of the factual correctness of the generated conclusions by combining extractive and abstractive learning objectives (Liang et al., 2022),
- evaluation of the performances of the presented approaches by a human expert (Liang et al., 2022).

1.4.3 Assessing distributional shifts for text classification

Radiology reports contain longitudinal information on a patient’s disease status, which is crucial for decision-making and outcome estimation (Fink et al., 2022a,b; Yim et al., 2016). Due to the continued high use of the narrative text form in radiology reports, the automated extraction of timelines and key clinical end-points, such as response to therapy and disease progression, is an essential subject of research in NLP (Fink et al., 2022a; Kehl et al., 2019; Agaronnik et al., 2020; Banerjee et al., 2019). Well-annotated data is needed to train and evaluate an AI system for automated information extraction. However, as mentioned earlier, only limited available well-annotated data sources exist because their creation is costly and time-consuming (Willemink et al., 2020). At the same time, SORs find their way into clinic practice (Fink et al., 2022a; ESR, 2018; Nobel et al., 2020; Weber et al., 2020). The growing resources of SORs in combination with the vast amounts of free-text-oncology reports (FTORs) raises the question of whether information raised in SORs could be used as a reference to build AI systems for the automated extraction from FTORs. Therefore, the Sections 2.3, 4.3, 5.3, and 6.3, which build upon the publication "Deep Learning-based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports (Fink et al., 2022a)", try to exploit the data mining advantages of SORs to train a NLP model for classifying tumor response categorys (TRCs) in FTORs without prior domain-specific feature engineering (Fink et al., 2022a). The main contributions include the following:

- the creation of a human reference annotation for the used FTORs and of human baseline performances by readers with varying levels of radiologic expertise. While doing so, the general question of how difficult it is to assign a TRC to a report is also examined,

- the creation of machine-based baseline performances on the text classification task. Besides a transformer-based implementation using the German BERT model, a traditional TF-IDF-based method using three different classifiers is implemented. One major challenge for the algorithms is to handle the distributional shift between structured training data and free-text test data,

- an in-depth analysis and discussion of the properties and characteristics of the used datasets and their influence on human and machine-based performances. Questions of interpretability and explainability of the results are also qualitatively addressed on a small scope.

The primary motivation to include the evaluation of traditional TF-IDF-based text classification methods is because the three datasets used in the experiments are from different institutions and contain domain-specific vocabulary. Especially for German clinical documents, it is still a subject of research how transformer-based models adapt and perform on data with a slight distributional shift and unknown vocabulary. In addition, (Chen et al., 2019) points out that conventional methods are still competitive with DL methods in the clinical domain.

1.4.4 Manual data annotation and tagging for medical images

Willemink et al. (2020) describe the curation and annotation of data as well as the availability of sufficient data, as substantial barriers for developing well-generalizing AI algorithms. For medical images, data acquisition involves ethical approval and data access, querying and de-identifying data, creating a patient cohort, and finally annotating the images (Willemink et al., 2020). In the best case, the metadata of DICOM images is sufficient to automatically create a cohort for training and evaluating an algorithm (Gauriau et al., 2020). However, often, manual curation steps while viewing the images are necessary. In addition, inconsistent or wrong metadata might hamper the automated cohort definition and reusability of DICOMs. Besides data curation, image classification and image-text correlation algorithms also rely on tags or free-text-based annotation. Other MIC algorithms require pixel-wise annotations, which tools like the Medical Imaging Interaction Toolkit and others already provide (Nolden et al., 2013; Wolf et al., 2004; Rubin et al., 2008). Because most existing software tools focus on pixel-wise annotation or are limited to non-DICOM data, the Sections 2.4, 4.4, 5.4, and 6.4 focus on whole image-based annotations. They are based on the publication

"Efficient DICOM-Image Tagging and Cohort Curation within Kaapana (Kades et al., 2022a)". The main contributions include the following:

- presentation of a technical solution for fast and intuitive DICOM image tagging by combing the software tools Doccano, OHIF and Kaapana,

- evaluation of the implementation by the feasibility of three use cases.

The use cases include the correction of wrong metadata in DICOM images, the addition of free-text information to DICOM images, and the possibility to label DICOM images for ML and DL applications (Kades et al., 2022a).

1.4.5 Real-world federated learning for the task of image segmentation

Providing a technical setup for federated scenarios is crucial, on the one hand, to increase the amount of available data and, on the other hand, to develop robust and generalizable AI algorithms (Willemink et al., 2020; Rieke et al., 2020). Especially to fight pandemics like COVID-19, a hospital-overarching application of AI algorithms is essential (Ting et al., 2020). One goal of RACOON is to automate and standardize the assessment of COVID-19-related tissue alterations through the training of segmentation algorithms and thus the extraction of biomarkers. As a first step, a collective of hospital-overarching patients was analyzed using structured reporting within RACOON. In addition, COVID-19 related anatomies and pathologies were segmented in the lung for each case. A next step would be the training of a segmentation algorithm on the segmented data that is robust and generalizes well on unseen data. For example, when applied to out-of-distribution data, the image segmentation algorithm nnU-Net has been shown to perform poorly (Isensee et al., 2021; Gonzalez et al., 2021; Full et al., 2021). Federated approaches might be a promising solution to train a generic and robust model that works well on data from all participating hospitals. However, enabling federated learning strategies across multiple clinical sites comes with challenges. On the one hand, these include a standardized infrastructure and execution environment that comply with data privacy and security and provide the technical possibility for federated learning strategies (Xu et al., 2021; Rieke et al., 2020). On the other hand, the used algorithms need to be able to handle data heterogeneities during training, and after training, the algorithms must perform well on in-house and unseen data from other clinics. As the Kaapana-based infrastructure used within RACOON lacks those requirements, the Sections 2.5, 4.5, 5.5, and 6.5, based on the publication "Towards Real-World Federated Learning in Medical Image Analysis Using Kaapana (Kades et al., 2022b)", present implementations of functionalities that allow federated use cases in the Kaapana toolkit, tailored to meet the technical and political requirements within RACOON. The main contributions include the following:

- implementation of a backend and a user interface as well as adjustments to local data processing pipelines within Kaapana to facilitate federated learning,
- an adaptation of the image segmentation algorithm nnU-Net to be used for federated learning,
- the creation of a baseline performance of a federated trained image segmentation model against single-site, centralized and ensemble-trained models on a multi-site prostate segmentation dataset (Liu et al., 2020a,b), with data collated from six different clinical sites.

Experiments that include training and testing on data from the same or different clinical sites help to assess whether the nnU-Net is suitable for the application in RACOON and which training strategy is the best.

2 Related work

The introduction covered several challenges for transitioning AI algorithms into the clinic. This section discusses related works to the five problems presented in the outline.

2.1 Semantic modeling for semantic textual similarity

STS is a subtask of NLP that tries to determine the semantic similarity between two text snippets. Most of the non-domain specific STS are benchmarked and developed on the STS benchmark dataset (STSbenchmark, 2023), which comprises a collection of sentence pairs from the SemEval STS tasks from 2012 to 2017 (Cer et al., 2017). The General Language Understanding Evaluation (GLUE) General Language Understanding Evaluation (GLUE Benchmark, 2023) dataset also comprises the STS benchmark. The current leading method from Jiang et al. (2020) proposes a learning framework that makes the fine-tuning of pre-trained models more robust and efficient and results in more generalizable fine-tuned models. In detail, they introduce smoothness-inducing regularization and a Bregman proximal point optimization into their training. Furthermore, Wang et al. (2019) introduces new auxiliary tasks during pretraining to leverage better the language structure at the word and sentence level. Jeyaraj and Kasthurirathna (2021) presents a multi-layered semantic similarity network that incorporates different similarity measures based on network science, and Raffel et al. (2020) leverages transfer learning techniques. Also, in this work, a method combining BERT with different semantic similarity measures is introduced.

In the clinical domain, methods are primarily developed and benchmarked on the data from track 1, N2C2/OHNLP Track on Clinical Semantic Textual Similarity, of the 2019 National NLP Clinical Challenges (N2C2). The winner of the 2019th challenge leverages multi-task learning as a training approach, in which they iteratively fine-tune on different datasets Mahajan et al. (2020). The presented system from Yang et al. (2020) report the performance of multiple transformer-based models and their ensembles. They show that when pretraining the model first on a STS general corpus before fine-tuning it on the clinical dataset, the RoBERTa-large model generates the best performance (Liu et al., 2019c). Moreover, Li et al. (2021a) employs a text data augmentation method and a self-ensemble ALBERT model under semi-supervised learning (Lan et al., 2019). Finally, Chang et al. (2021) incorporates domain knowledge into the language model using graph convolutional networks.

To summarize, all state-of-the-art STS algorithms build on top of the transformer architectures and incorporate additional knowledge via transfer learning, multi-task learning, ensembling methods, graph neural networks, or other similarity measures. The introduced methods in this work follow similar design patterns.

2.2 Semi-structured data analysis for text summarization

For text summarization, mostly extractive and abstractive techniques, or a combination of both methods is often applied. In extractive methods, the text summarization consists of only salient text snippets from the source text. Often, human-designed features or rules define which parts are extracted. In abstractive approaches, a concise summary is generated based on a learned feature representation of the source text. While extractive approaches can be rule-based, abstractive approaches require DL architectures. However, also extractive models can be enhanced by using DL architectures. For example, (Nallapati et al., 2016) propose to use a RNN architecture to determine which sentences should be part of the final summary. Like machine translation models, abstractive text summarization algorithms are sequence to sequence (seq2seq) models with an encoder-decoder architecture and e.g. attentive RNNs (Nallapati et al., 2016; Nallapati et al., 2016; Chopra et al., 2016; Liang, 2021; Liang et al., 2022). By generating the summary word by word, abstractive models are more prone to generate syntactically and semantically incorrect summarization than extractive approaches. Also, abstractive methods often struggle with factual correctness and repeated contents (See et al., 2017). Therefore, extractive and abstractive approaches are often combined. For example, See et al. (2017) present a hybrid pointer-generator network that copies words from the source text to the target summary via pointing Vinyals et al. (2015). In detail, a generation probability decides whether to generate a word from given vocabulary distribution or copy a word from the input sequence (See et al., 2017). Other approaches start by extracting salient sentences from the source document before they use abstractive techniques to generate the summary based on the selected sentences (Chen and Bansal, 2018; Kryściński et al., 2018). This work combines the pointer network with a transformer-based encode-decoder model.

The field of automatic text summarization algorithms advanced with the introduction of pre-trained language models (Miller, 2019; Liu and Lapata, 2019; Zhang et al., 2019; Rothe et al., 2020). As for many other NLP downstream tasks, pre-trained language models such as BERT (Devlin et al., 2019) or Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019) can be fine-tuned for text summarization tasks. There exist many different seq2seq models such as BERT2Random, BERT2BERT, BERT2GPT. The semantic knowledge already incorporated into a pre-trained language model reduces the computational resources and needed time, while improving the performance of the sequence generation (Liang et al., 2022; Rothe et al., 2020). This work fine-tunes a pre-trained German BERT model for summarizing German radiology findings using abstractive and extractive text summarization techniques.

To summarize clinical reports, Nguyen et al. (2020) have worked on summarizing Chest X-ray reports with LSTM models and an encoder-decoder model incorporating attention. Zhang et al. (2018) were one of the first to apply pointer-generator networks in combination with LSTM networks to summarize radiology reports. In addition, Zhang et al. (2018) encoded the background information and findings in two different encoders. In contrast to Zhang et al. (2018), in this work, the pointer-generator network is combined with a transformer-based encoder-decoder, and only one encoder for the background and findings information is used. A later proposed model from Zhang et al. (2020) incorporated a reinforcement learning approach to improve factual correctness. Works leveraging ontologies include Sotudeh Gharebagh et al. (2020), which improved the clinical abstractive summarization by augmenting salient ontological terms into the summarizer. Furthermore, MacAvaney et al. (2019) introduced an ontology-aware pointer-generator network to leverage domain-specific knowledge encoded in an ontology to improve the summarization. Similar to Zhang et al. (2020), Sotudeh Gharebagh et al. (2020) and Nguyen et al. (2020), the generated summaries in this work are evaluated by humans in terms of factual correctness and fluency. Besides summarizations based on background and findings information, Lovelace and Mortazavi (2020), Li et al. (2018b) and Liu et al. (2019) present approaches to generate a Chest-X ray radiology report directly from the medical image.

2.3 Assessing distributional shifts for text classification

Before the introduction of DNNs, rule-based approaches in combination with machine learning approaches were used for text classification tasks. In detail, features are created from the text using BOW, TF-IDF or a purely heuristic approach. The features are then fed into a classification algorithm like Naïve Bayes, SVM, random forests, etc. (Minaee et al., 2021). With NNs, models evolved that create an embedding representation of an input sequence such as Latent Semantic Analysis (LSA). Furthermore, pre-trained embeddings like word2vec (Mikolov et al., 2013) evolved with better computational resources and data availability. The embedding representations serve as input for a classification layer in text classification tasks. Current state-of-the-art text classification algorithms use transformer-based architectures (Minaee et al., 2021; Vaswani et al., 2017). Transformer-based models are used for many different NLP downstream tasks. In text classification, pre-trained models such as BERT are fine-tuned using a so-called classifier token in combination with a classification layer.

Multiple challenges occur when using transformer-based models on clinical datasets. Those problems include the domain shift in comparison to public data, very long sequences, imbalanced classes, and interpretability. Often models are further pre-trained on domain-specific datasets such as clinical notes to incorporate domain-specific knowledge into a model (Alsentzer et al., 2019; Lee et al., 2020). To handle long texts, many approaches suggest chunking long sequences as input for transformer-based models. Wu et al. (2021) proposes to process the classification token of several chunks by feeding their values into an LSTM or another transformer model. Mulyar et al. (2019) propose to take the mean or a concatenation of the classification tokens and feed the result into a classification layer for the phenotyping of clinical notes. To process only the task-relevant chunks from long clinical tests, Huang et al. (2020) proposes to pre-select text snippets using a regex expression. Furthermore, instead of chunking long sequences, new models were designed to overcome the memory consumption of $\mathcal{O}(n^2)$ of the original BERT model. Beltagy et al. (2020), Zaheer et al. (2020), and Ainslie et al. (2020) introduced new attention patterns, which combine local attention with global attention. Instead of full attention (n to n attention of each token), they use a mixture of dilated and sliding window attention to capturing local context. They introduce sparse and random attention or attention to a few pre-selected input locations for global context. The present work follows the chunking approach from Mulyar et al. (2019) to process long sequences.

The interpretability of transformer models is an area of ongoing research and of particular importance in the medical domain to understand the machine's reasoning when used in clinical applications. The role and importance of attention of fine-tuned models have been evaluated by Kovaleva et al. (2019); Kobayashi et al. (2020); Sun and Lu (2020); Hao et al. (2020). Whether attention can be used for explanation and different tests to assess explainability via attention, have been discussed in Jain and Wallace (2019); Serrano and Smith (2019); Vashishth et al. (2019); Wiegreffe and Pinter (2019). For example, Serrano and Smith (2019) observed that attention does not necessarily reflect importance after manipulating attention weights in already trained text classification models. Vashishth et al. (2019) show that attention is only interpretable for natural language processing tasks other than text classification. By using a hierarchical approach, Huang et al. (2020) claims to improve the interoperability of the transformer-based model after fine-tuning on clinical notes. This work will qualitatively show the attention weights of a fine-tuned BERT model and the weights of a TF-IDF analysis.

Also, in text classification tasks on free-text radiology reports, transformer-based models have shown superior performance compared to feature-based methods (Steinkamp et al., 2019; Linna and Kahn, 2022; Fink et al., 2022a; Gehrmann et al., 2018). In oncology, many works focus on extracting timelines and key clinical endpoints, such as response therapy and disease progression (Fink et al., 2022a; Kehl et al., 2019; Agaronnik et al., 2020; Banerjee et al., 2019; Pons et al., 2016; Smit et al., 2020). To improve the text classification performance of BERT on radiology reports, Smit et al. (2020) proposes to fine-tune the model first on automatically generated labels and then on the expert-level annotations. In this work, automated annotations from SORs are leveraged to train text classification algorithms for the application on FTORs.

2.4 Manual data annotation and tagging for medical images

Efficient cohort selection, data curation, and data annotation tools are crucial to generating high-quality labeled datasets for the applications of AI algorithms (Willemink et al., 2020). For medical images, the DICOM metadata provides valuable information on the underlying image. Diaz et al. (2021) provide a comprehensive guide of steps that are necessary to prepare medical images for the development and application of AI algorithms. Gauriau et al. (2020) successfully used the DICOM metadata to automate the identification of brain MRI sequences. However, human-made metadata might be incorrect (Magudia et al., 2021; Schuhegger, 2021), which hampers automation to assemble large medical image datasets. Almeida et al. (2021) proposed to combine the information from EHR and DICOM metadata to define cohorts over medical imaging datasets. This work presents an interactive tool for dedicated, workflow-integrated, and standard-conform manual examination of medical images belonging to a cohort and the functionality to curate wrong DICOM metadata (Kades et al., 2022a). After an effective cohort selection and data curation, the annotation of data is the next crucial step. While there exist many tools for the creation of pixel-wise annotations of DICOM images which include the MITK (Wolf et al., 2004; Nolden et al., 2013), 3D Slicer (Lasso, 2019), DicomAnnotator (Dong et al., 2020) or iPad (Rubin et al., 2008) and ePad (Rubin et al., 2014), tools to generate image-level annotations of DICOM images are still sparse. For this reason, the presented software tool also integrates a fast and intuitive image-level tagging of DICOM images (Kades et al., 2022a). The tagging functionality can also be used as a post-processing step to evaluate the results of an AI algorithms. (Stein et al., 2019) presented a similar tool specifically designed to efficiently evaluate automatically generated segmentations based on different metrics and visualizations.

2.5 Real-world federated learning for the task of image segmentation

The research in the field of Federated Learning (FL) is diverse, ranging from purely methodical to technical, software-related contributions. Purely methodical contributions focus on algorithmic solutions to handle not independent and identically distributed (Non-IID) data and improve communication efficiency while reaching comparable performance as centralized trained models (Li et al., 2020; Kairouz et al., 2021). Furthermore, security-related methods in FL address data protection issues and privacy-preserving methods, e.g. differential privacy and homomorphic encryption methods. The more software-related research areas deal with the technical implementation of FL systems, which include finding solutions for exchanging model weights and for handling failures, such as network failures, in FL systems (Rieke et al., 2020; Xu et al., 2021). In the technical realization, security aspects are also of high importance.

From a technical point of view, (Li et al., 2021b) provides a systematic overview of domain-independent open-source software solutions for FL along with different FL averaging strategies. Presented solutions include FATE (FATE community, 2023), TensorFlow Federated (TensorFlow, 2023), OpenMined (OpenMined, 2023), PaddleFl (PaddlePaddle, 2023), or FedML (FedML, 2023). All the tools have advantages and disadvantages depending on the use cases. In the medical domain, Kaissis et al. (2020) presents the free open-source framework Privacy-preserving Medical Image Analysis (PriMIA) with the support of differential privacy and secure aggregation in federated learning and encrypted inference on medical imaging data. PriMIA was successfully applied in a real-life case study to classify pediatric chest X-rays automatically. Another tool for federated learning in real-world clinical settings is the proprietary healthcare application framework Computational platform to build manage and deploy intelligent medical imaging workflows and instruments (NVIDIA Clara), which features federated learning functionalities in combination with the NVIDIA Clara Train SDK. The software stack from NVIDIA was successfully applied in real-world scenarios (Yang et al., 2021; Dayan et al., 2021; Roth et al., 2020; Sarma et al., 2021; Roth et al., 2021). In conjunction with Medical Open Network for Artificial Intelligence (MONAI) further efforts to enable federated learning in the medial imaging domain are pursued by OpenFL, Flower and Substra. The technical solution for federated learning presented in this work follows technical prerequisites given by RACOON. A key requirement is a uni-directional communication from the client instances to the central instance. In addition, many algorithms already run locally within RACOON and should be extended with a federated functionality. Therefore, the presented implementation here tries to be as agnostic as possible by wrapping the federated functionality around existing workflows without the need to customize them on a code level.

From a methodical point of view, Rieke et al. (2020) and Prayitno et al. (2021) summarized and assessed the challenges of federated learning in clinical environments . Ziller et al. (2021) and Kaissis et al. (2020) discuss secure and privacy-preserving methods such as differential privacy. Sheller et al. (2020) and Rieke et al. (2020) present and evaluate different federated learning strategies versus centralized training. Furthermore, multiple works compare models trained locally against models trained across multiple institutions in a federation (Dayan et al., 2021; Roth et al., 2021, 2020; Sarma et al., 2021). This work follows one of the first federated learning strategies, the $FedAvg$ approach introduced by Brendan McMahan et al. (2016). Many works to benchmark federated learning are based on publicly available datasets such as the BraTS datasets (Menze et al., 2015) or the multi-site prostate segmentation dataset (Liu et al., 2020a) presented in Section 3.2.3 (Sheller et al., 2020; Gu et al., 2021a; Jiang et al., 2022; Liu et al., 2020a,b, 2021). Furthermore, many works using federated learning in conjunction with COVID-19 exist (Dayan et al., 2021; Yang et al., 2021; Dou et al., 2021; Feki et al., 2021).
To evaluate the performance of different training strategies against generalizability and robustness, the multi-site prostate segmentation dataset presented by (Liu et al., 2020a) serves as a benchmark dataset in this work. Domain generalization and cross-domain performance on the multi-site prostate segmentation dataset were, amongst others, assessed in (Liu et al., 2021; Jiang et al., 2022). Research to evaluate the generalizability of the nnU-Net, used in this work as the segmentation algorithms, to data from unseen institutions is still sparse. However, e.g. Full et al. (2021) study the generalization and robustness of the nnU-Net on a dataset of cardiac MRI as well as suggest ways to improve the generalizability of the nnU-Net. Finally, Gonzalez et al. (2021) presents a method to detect when pre-trained nnU-Net models fail.

3 Materials

3.1 Kaapana as an open-source toolkit for provisioning medical imaging platforms

Working in multi-institutional settings requires a technology that satisfies multiple requirements. It has to run locally inside a clinical infrastructure, it must be able to receive, store and index clinical data like imaging and text data and it should allow for standardized execution of algorithms on a selected subset of data (cohort). In this work, the Kaapana open-source platform provision toolkit (Scherer et al., 2023) is used as the base technology upon which additional features required to execute experiments are implemented. The following section describes the main building blocks of the Kaapana technology stack. It is mainly based on the work "Joint Imaging Platform for Federated Clinical Data Analytics" (Scherer et al., 2020) and the thesis "Decentralized Infrastructure for Medical Image Analysis" (Scherer, 2022).

3.1.1 The technological foundation of Kaapana

Kaapana itself is a toolkit to build platforms for medical image analysis that run on-premise as a private cloud computing setup inside a local clinical IT landscape. From a technical point of view, a traditional private cloud computing setup consists of three main building blocks, an infrastructure (Infrastructure-as-a-Service (IaaS)), a so-called platform (Platform-as-a-Service (PaaS)) and an application (Software-as-a-Service (SaaS)) layer (Scherer et al., 2020; Mell and Grance, 2011). In the case of Kaapana, Scherer et al. (2020) added an additional container virtualization

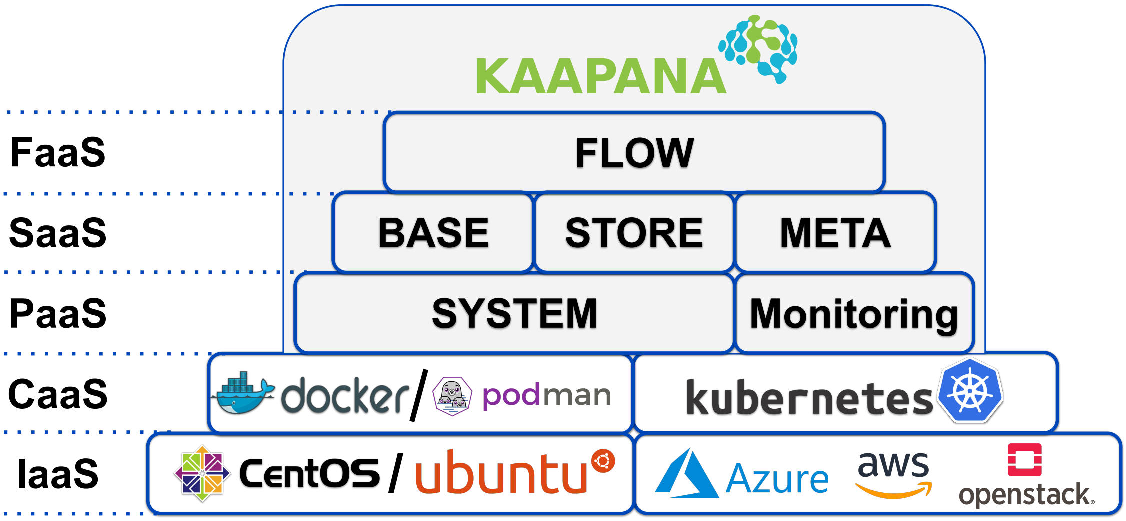


Figure 3.1: *Illustration of the private cloud computing setup of Kaapana consisting of the different service layers. This figure was adapted from Scherer (2022). Kaapana=Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis.*

layer (Containers-as-a-Service (CaaS)) and a function layer (Functions-as-a-Service (FaaS)) for on-demand data processing (Scherer et al., 2020; CNCF, 2023). Figure 3.1 from (Scherer, 2022) illustrates the five layers, containing also the five main functional units "SYSTEM", "MONITORING", "BASE", "STORE", "META" and "FLOW" which are introduced to group components of similar functions in Kaapana.

The IaaS and CaaS layer

In the IaaS layer, Kaapana currently supports Ubuntu, CentOS or AlmaLinux as an operating system (OS). It can be deployed on common public IaaS providers like Amazon Web Services (AWS) (Amazon, 2023b) or on Microsoft Azure (Microsoft, 2023) as well as on-premise on dedicated private hardware, which is the default when deployed in clinics (Scherer, 2022; Scherer et al., 2020). On the CaaS layer, Kaapana uses Kubernetes (K8s) (Rensin, 2015) as a container orchestration system. Therefore, all services and processing pipelines of the platform run within containers on the Kubernetes cluster. This architecture provides flexibility in managing the container life cycle and enables scaling and monitoring of individual containers. Furthermore, since containers provide their own execution environment including the OS layer, a wide variety of software can be supported and therefore easily integrated into Kaapana. The containers of Kaapana are built using the container engines docker (Merkel, 2014) or

Podman (Podman community, 2023). To manage the resources needed for a platform deployment, Kaapana uses Helm (The Linux Foundation, 2023), which is specifically designed to package and manage Kubernetes configuration files. To deploy Kaapana on an OS, firstly, with the help of an installer script, necessary dependencies like drivers along with the Kubernetes cluster using Microk8s (Microk8s, 2023) are set up. Secondly, a Helm chart is installed containing all necessary Kubernetes configurations for the Kaapana platform. Both, the needed containers as well as the needed helm charts are either retrieved from an Open Container Initiative (OCI) -based registry or are uploaded locally directly into the Kubernetes cluster (Scherer, 2022; Scherer et al., 2020).

The PaaS layer - SYSTEM and MONITORING units

The platform layer contains the base components needed for a minimal version of a platform. It features interfaces for the provisioning of applications and micro-services on the underlying Kubernetes cluster. The core of Kaapana consists of a reverse proxy, an authentication proxy and an authentication provider. As a reverse proxy, Kaapana uses the free open-source components of Traefik (Ludovic Fernandez, 2023) along with webOAuth2Proxy2023 (OAuth2 Proxy, 2023) as an authentication proxy. All user interfaces provided in Kaapana are accessible from a single point of entry on Hypertext Transfer Protocol Secure (HTTPS). The open-source tool Keycloak (Keycloak, 2023), a central OpenID Connect (OIDC) enabled authentication provider, is used as an identity and access management solution to support single sign-on (SSO). The core components of Kaapana are completed by a monitoring system based on Prometheus (Prometheus, 2023b) as an event monitoring and alerting tool, of Alertmanager (Prometheus, 2023a) to forward alerts sent by the Prometheus server and of Grafana (Grafana, 2023) to visualize the gathered metric on a dashboard (Scherer, 2022; Scherer et al., 2020).

The SaaS layer - BASE, STORE and META units

The software-as-a-service layer facilitates the customization of the platform. In its default configuration, Kaapana adds three functional units BASE, STORE and META. In BASE, a central web application developed with the frontend framework Vue.js (Vue.js, 2023) and the design framework Vuetify (Vuetify , 2023) is hosted. The user interface, denoted as a "landing page", facilitates access to all services that run on different sub-paths by merging them with the help of inline frame elements (iFrames) into one single-page application (SPA). In addition, the landing page includes the management of so-called extensions. Extensions are essentially a set of micro-services that are defined in a Helm chart, which can be installed and uninstalled via the management system on the landing page (Scherer, 2022; Scherer et al., 2020). In the functional unit STORE, all services related to data storage, access and visualization are located. As described in Section 1.2.1 DICOM is the primary format for the storage and communication of medical imaging data. The Kaapana platform offers a C-STORE (NEMA, 2023) and a DICOM web (Genereaux et al., 2018) interface to receive data directly from the clinical PACS (Scherer, 2022). The platform runs an internal research PACS via the Open Source Clinical Image and Object Management application dcm4chee (Gunter Zeilinger, 2021) in which upon arrival all images are stored. In addition, Kaapana integrates the zero-footprint web-based OHIF Medical Imaging Viewer to visualize the stored DICOM data in the browser (Ziegler et al., 2020). The images can be selected and viewed based on the Study ID attribute from the metadata of the DICOM images are directly retrieved from the internal PACS. The OHIF viewer enables besides the visualization of three-dimensional scans such as CT or MRI and corresponding annotations, also several image manipulation and annotation tasks (Scherer, 2022). In order to also store data other than medical imaging data, Kaapana provides the High Performance Object Storage MinIO (MinIO, 2023), which is compatible with the widely used Amazon Simple Storage Service (S3) API (Amazon, 2023a; Scherer, 2022). Since it is crucial for a medical imaging platform to have an overview as well as to filter and query data that is stored on the platform, Kaapana is equipped with Elasticsearch as a search engine and Kibana as a data visualization dashboard. All relevant metadata of incoming DICOM images is stored in the Elasticsearch database. Different dashboards are provided for viewing, querying and filtering the metadata of the images depending on the task at hand. Besides providing an overview of the data, the main purpose of Kaapanas' metadata capabilities is to allow a fine-grained definition of cohorts via search queries that later can be processed via processing pipelines. The Kibana dashboard is equipped with a custom plugin that makes it easy to parameterize and trigger processing pipelines. When the cohort is selected in the Kibana dashboard, the plugin presents a form to parameterize the processing pipeline. When the start button is pressed, a request containing the search query from the dashboard and the parameterization from the plugin form is sent via an HTTP request to the executing unit described in the following section.

The FaaS layer - FLOW unit

Next to data storage, viewing, and querying the main functionality of Kaapana is the execution of processing pipelines, which is why the Function-as-a-service layer is introduced for the on-demand scheduling of processing jobs. In contrast to the services of the PaaS and the Saas layer, the services running in the FaaS layer run only for a time until their specified job or function is fulfilled. In Kaapana, the functional unit FLOW consists of the workflow management system Apache Airflow (Apache, 2023). Airflow manages, schedules and monitors on-demand jobs and services which form together a data processing pipeline. Throughout the work, the terms "processing pipeline" and "workflow" have the same meaning and refer always to a Directed Acyclic Graph (DAG) representing the tasks a workflow carries out and their interdependence. For the actual execution of processing pipelines, Kaapana extends Airflow with a plugin that enables the execution of a workflow via a REST API request. Within Kaapana this request is generally triggered from the Kibana dashboard plugin, as mentioned in the last section. The requests contain along with the search query also pipeline specific parameters, which are defined in the DAG. In Airflow a DAG consists of Airflow operators, which are essentially jobs that are executed in a specific order. An operator in its most basic form executes a Python script. However, Kaapana extended the operators concept to be able to dynamically launch containers on the Kubernetes Cluster. Therefore, an operator can execute any software which can be packed into a container, independent of the desired OS or additional dependencies.

3.1.2 Deployment and use cases of Kaapana

Kaapana deployment within the DKTK and RACOON

Section 1.2.3 introduces two initiatives using Kaapana, which are the JIP initiated by DKTK and RACOON. Within the DKTK-sites the platform is deployed on a dedicated hardware server inside the clinical IT landscape. Figure 3.2 illustrates the setup at the hospitals in the DKTK. The container images needed for the platform installation are retrieved from a container registry located at and controlled by the DKFZ. This ensures that only verified container images are distributed to the clinics. The platform is only accessible from within the clinical network infrastructure. In a typical use case scenario, a user sends imaging data from the clinical PACS to the platform, accesses the platform from their workstation via a web browser, defines their cohort on the Kibana dashboard and triggers a processing pipeline on Airflow. If the executed DAG generates results, they are either send to the internal PACS or to the object store High Performance Object Storage (MinIO).

Figure 3.3 illustrates the technical infrastructure as it is deployed in RACOON. The RACOON Node corresponds to the unit, which is set up on dedicated hardware inside each participating hospital. Each node runs Windows Server (Microsoft, 2023), which provides a Hyper-V virtualization layer. Respective applications of the three sections RACOON REPORTING, RACOON JIP and RACOON SATORI are all deployed on Hyper-V. RACOON REPORTING consists of a certified medical product, Mint Medical (Mint Medical, 2023), which meets all requirements for clinical use. It is the only entry point for data from within the hospital. Furthermore, it enables structured reporting for COVID-19 patients and stores all data in a consistent way (Scherer, 2022). All data transferred to RACOON JIP or RACOON SATORI are anonymized to ensure a high level of data privacy. Satori (Fraunhofer MEVIS, 2023) is a dedicated research tool developed by MEVIS for efficient pixel-wise annotation of three-dimensional

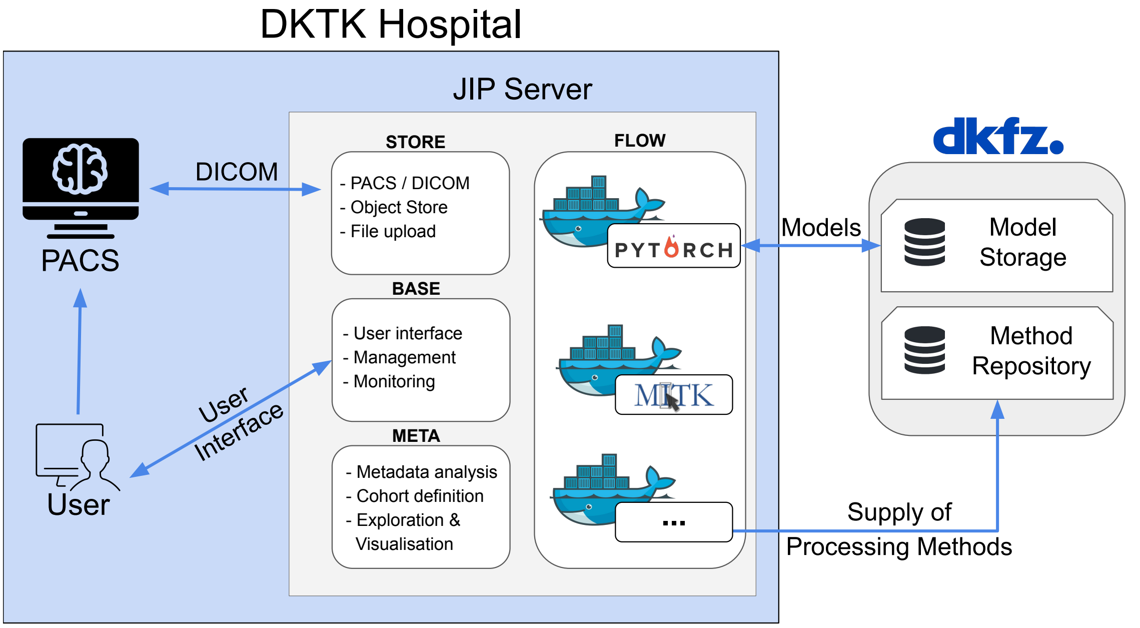


Figure 3.2: *Setup of the JIP within a clinical center. The platform itself is running on dedicated hardware within the clinical network. The user is able to send DICOM images to the platform and to interact with the platform via browser access. The platform is connected to a central container registry to retrieve necessary container images for its services and processing pipelines. This figure was adapted from Scherer (2022). DICOM=Digital Imaging and Communications in Medicine, JIP=Joint Imaging Platform, PACS=Picture Archiving and Communication System.*

images (Scherer, 2022). The setup of RACOON JIP is similar to the one used in DKTK, except that it is deployed on a virtual machine, and support for complete offline installation without an external container registry was introduced.

nnU-Net training as use case

Kaapana can be used in many different scenarios. On one side, it can be used to deploy software by providing an application as an extension to the platform. On the other side, it offers the possibility to apply processing pipelines on a specific set of data. In this work, the nnU-Net Isensee et al. (2021) integrated within Kaapana is a central segmentation method, this section will only cover the nnU-Net training and evaluation workflow. More use cases for Kaapana can be found in (Scherer, 2022).

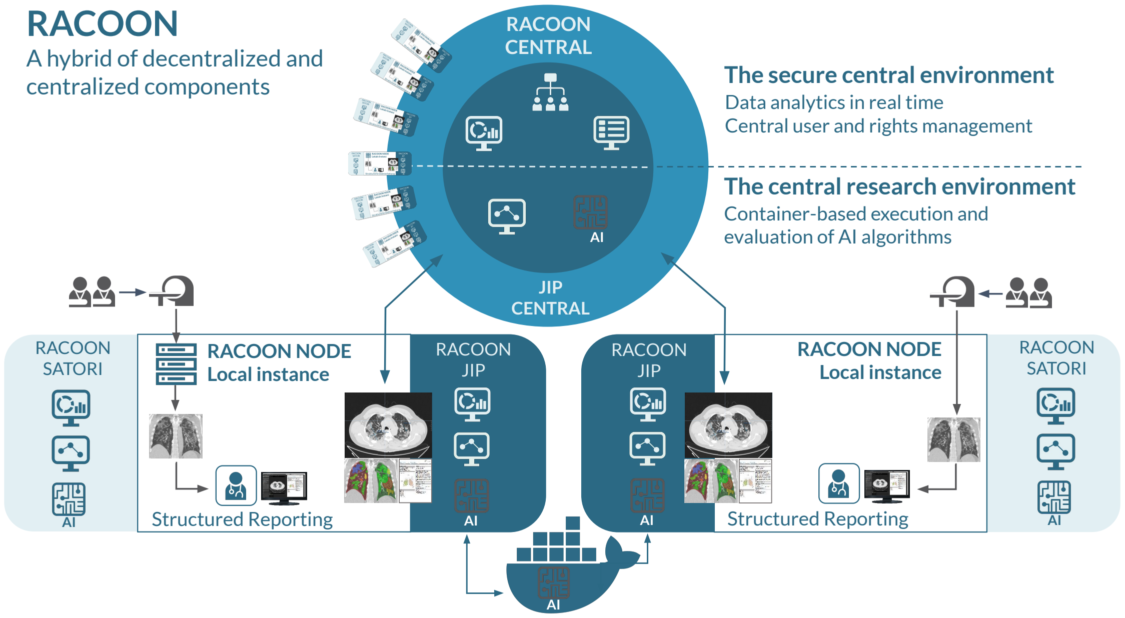

RACOON
A hybrid of decentralized and centralized components
RACOON CENTRAL
The secure central environment
Data analytics in real time
Central user and rights management
The central research environment
Container-based execution and evaluation of AI algorithms
AI
JIP CENTRAL
RACOON SATORI
AI
RACOON NODE
Local instance
Structured Reporting
RACOON JIP
AI
RACOON JIP
AI
RACOON NODE
Local instance
Structured Reporting
RACOON SATORI
AI
AI

Figure 3.3: *Infrastructure stack as it is deployed within RACOON. A JIP instance adapted for RACOON runs in a central instance as well as in all participating clinical centers. Alongside the RACOON JIP, Mint Medical and Satori are deployed. The RACOON JIP receives medical data from the Mint Medical system and its main task is the application of AI methods on the received data. This figure was adapted from Scherer (2022). AI=artificial intelligence, JIP=Joint Imaging Platform, RACOON=Radiological Cooperative Network.*

nnU-Net training workflow

The nnU-Net is a DL-based segmentation method that automatically configures the preprocessing, the network architecture, the training and the post-processing based on an introduced recipe. This recipe is based on a set of fixed, rule-based and empirical parameters. Fixed parameters are predefined and independent of the training dataset. They include e.g. the learning rate, the number of epochs or the loss function. The rule-based parameters are determined by interdependent heuristic rules based on a "dataset fingerprint", which contains information about the distribution of spacing, the median shape, the intensity distribution and the imaging modality of the training dataset. Examples of rule-based parameters are image target spacing, intensity normalization or patch size. The empirical parameters determine possible post-processing or ensemble selections. More information about the selection of hyperparameters is given in (Isensee et al., 2021).

How the processing pipeline of the nnU-Net training is implemented in Kaapana is

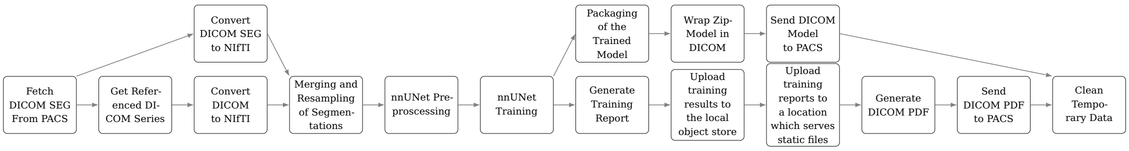

Fetch DICOM SEG From PACS
Convert DICOM SEG to NIfTI
Get Referenced DICOM Series
Convert DICOM to NIfTI
Merging and Resampling of Segmentations
nnUNet Preproscessing
nnUNet Training
Packaging of the Trained Model
Generate Training Report
Wrap Zip-Model in DICOM
Upload training results to the local object store
Send DICOM Model to PACS
Upload training reports to a location which serves static files
Generate DICOM PDF
Send DICOM PDF to PACS
Clean Temporary Data

Figure 3.4: *Detailed view of the processing pipeline for the nnU-Net training DAG and its corresponding operators. This figure was adapted from Scherer (2022). DICOM=Digital Imaging and Communications in Medicine, NIfTI=Neuroimaging Informatics Technology Initiative, nnU-Net=no new net U-Net, PACS=Picture Archiving and Communication System, SEG=Segmentation Objects.*

illustrated in Figure 3.4. The nnU-Net training DAG is composed of several operators. In the first step, the segmentations used for training are retrieved from the internal PACS, before they are converted to the NIfTI format (upper row). In parallel, the referenced source images are retrieved from the PACS and converted to the NIfTI format. In the next step, the source images and the segmentations are merged and resampled, so that for each source image only one single NIfTI file with consistent label encodings and without any mask overlaps exists (Scherer, 2022). The next operator first determines the preprocessing and network architecture and then executes the preprocessing. Then the actual nnU-Net training is executed in the following operator. The two operators are mainly based on *batchgenerators* (Isensee et al., 2020) and the nnU-Net implementation (Isensee et al., 2021; nnU-Net, 2023). In the following operators, a training report containing the training parameters along with training progress graphs is generated and uploaded together with the trained model to the local object storage MinIO. A DICOM-converted version of the model and report is uploaded to the internal PACS as well. The generated report is uploaded to a special location within the local object store that serves static files via the web interface of the platform, and that allows the examination of the report. In the final step, all temporary files of the pipeline are removed.

nnU-Net ensemble workflow

Once one or multiple nnU-Net models are trained they can be ensembled and evaluated using the nnU-Net ensemble workflow. In a nnU-Net ensemble, the softmax probabilities generated by selected models on target source images are averaged to create a single final prediction. In Figure 3.5 the nnU-Net ensemble implementation in Kaapana is illustrated. In the upper branch, all nnU-Net models needed for the ensembling, and evaluation are downloaded. In the lower branch, the testing cohort is prepared. Here, the reference segmentations along with the corresponding source images are retrieved from the internal PACS and converted to the NIfTI format. Then

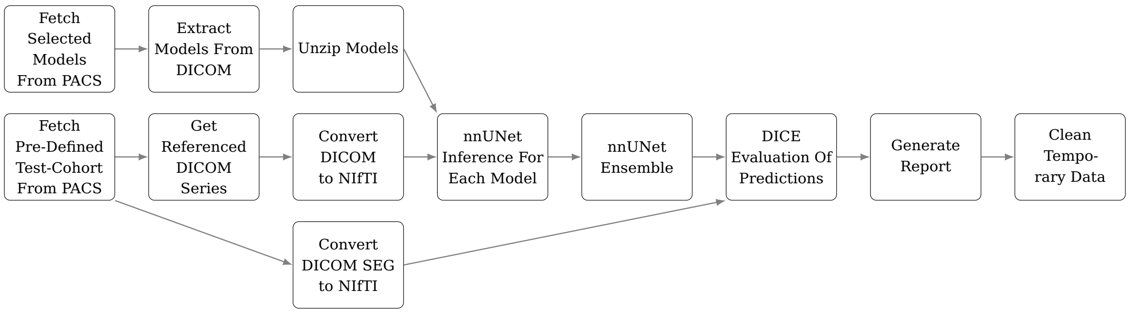

Fetch Selected Models From PACS
Extract Models From DICOM
Unzip Models
Fetch Pre-Defined Test-Cohort From PACS
Get Referenced DICOM Series
Convert DICOM to NIfTI
nnUNet Inference For Each Model
nnUNet Ensemble
DICE Evaluation Of Predictions
Generate Report
Clean Tempo-rary Data
Convert DICOM SEG to NIfTI

Figure 3.5: *Detailed view of the processing pipeline for the nnU-Net ensemble DAG and its corresponding operators. This figure was adapted from Scherer (2022). DICOM=Digital Imaging and Communications in Medicine, DICE=Sørensen-Dice coefficient, NIfTI=Neuroimaging Informatics Technology Initiative, nnU-Net=no new net U-Net, PACS=Picture Archiving and Communication System, SEG=Segmentation Objects.*

the actual inference and ensemble for all images of the testing cohort with all models are executed. After checking if all the created segmentations are valid and after removing invalid segmentations, the Sørensen-Dice coefficient (DICE) and the Average Surface Distance (ASD) scores for the predicted segmentations are calculated. Finally, a report containing the scores along with box plots is created and uploaded to the local object store in the location that servers static files for later review. In the last step, again all temporarily created files are removed. The workflows can also be used to only evaluate one model, then the ensembling-related steps are omitted.

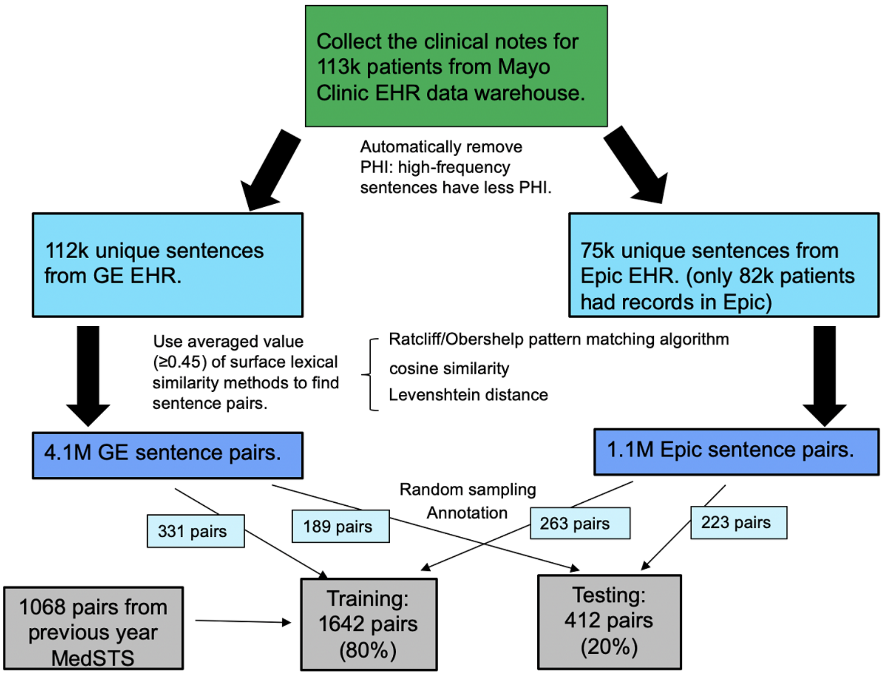

Collect the clinical notes for 113k patients from Mayo Clinic EHR data warehouse.
Automatically remove PHI: high-frequency sentences have less PHI.
112k unique sentences from GE EHR.
75k unique sentences from Epic EHR. (only 82k patients had records in Epic)
Use averaged value (≥0.45) of surface lexical similarity methods to find sentence pairs.
Ratcliff/Obershelp pattern matching algorithm
cosine similarity
Levenshtein distance
4.1M GE sentence pairs.
1.1M Epic sentence pairs.
Random sampling
Annotation
331 pairs
189 pairs
263 pairs
223 pairs
1068 pairs from previous year MedSTS
Training: 1642 pairs (80%)
Testing: 412 pairs (20%)

Figure 3.6: *Process of the creation of the dataset for the 2019 N2C2/OHNLP Clinical STS track. This figure was adapted from Wang et al. (2020). Clinical STS=Clinical Semantic Textual Similarity, EHR=Electronic health record, N2C2=National NLP Clinical Challenges, OHNLP=Open Health Natural Language Processing.*

3.2 Multi-institutional datasets

Three different datasets are used throughout the work. They are presented in detail in the following section.

3.2.1 Clinical Semantic Textual Similarity dataset

The Clinical Semantic Textual Similarity (Clinical STS) dataset consists of 2054 clinical sentence pairs, which were annotated by clinical experts with its degree of semantic textual similarity, i.e. to which degree two snippets of clinical text are semantically equivalent (Wang et al., 2018b,a, 2020). The dataset was created in the context of task 2 on Clinical Semantic Textual Similarity of the BioCreative/OHNLP challenge 2018 (Wang et al., 2018b,a; BioCreative/OHNLP Challenge 2018, 2023) and the 2019 N2C2/OHNLP Clinical STS track, which was part of the 2019 N2C2/OHNLP Shared-Task and Workshop (Wang et al., 2020; National NLP Clinical Challenges, 2023). The in total 2054 clinical sentence pairs of the 2019 N2C2/OHNLP Clinical STS track originated by combining 1068 clinical sentence pairs from the BioCreative/OHNLP ClinicalSTS shared task in 2018 with 1006 new sentence pairs. Duplicate clinical sentence pairs were removed in the process. All sentence pairs were picked from clinical notes of the Mayo Clinic EHR data warehouse. Figure 3.6 illustrates the process of how the sentence pairs were selected. After the selection, the sentence pairs were independently annotated by two clinical experts with many years of experience. In case of disagreement, the average score was taken as the reference standard. Therefore, the annotations contain integer and noninteger values ranging from 0 (not similar) to 5 (completely similar). The two annotators had a moderated agreement measured with a weighted Cohen's Kappa of 0.67 on the 1068 clinical sentence pairs and 0.6 for the additional 1006 sentence pairs. Examples of sentence pairs for each score are given in Table 3.7. The final dataset of the 2019 N2C2/OHNLP Clinical STS track consists of 1642 training and 412 test sentence pairs. More information on the challenge and the dataset can be found in (Wang et al., 2018b,a, 2020). In addition, in Section 5.1.2 additional analysis of the dataset is presented.

3.2.2 German radiology reports

The second dataset comprises a collection of radiology reports retrieved from the Radiology Information System (RIS) from three independent clinical sites, the University Hospital Heidelberg (UKHD), the German Cancer Research Center (DKFZ) and the Heidelberg Thoracic Clinic (TKH). All three radiology departments are associated with the German Cancer Research Center. The German radiology reports include all kinds of oncological diagnoses with slightly varying tumor entities depending on the clinical site and contain examinations of all body regions (Fink et al., 2022a). The retrospective studies comprising the German radiology reports are compliant with the Health Insurance Portability and Accountability Act and approved by the Institutional Review Board (S-083/2018). The requirements to obtain informed consent were waived. All anonymized German reports were stored locally on dedicated computing resources.

The collection of radiology reports consists of consecutive reports for CT, MRI and ultrasound (US) examinations of all body regions, acquired between March 2018 and August 2021. In total, 14569 reports were retrieved with 13685 SORs from the UKHD, which is a tertiary care center, 412 free-text-oncology reports from the German Cancer Research Center (FTOR-DKFZ) and another 472 free-text-oncology reports from the Heidelberg Thoracic Clinic (FTORT-TKH), which is a hospital specializing in chest diseases (Fink et al., 2022a).

The concept of structured oncology reporting was introduced by Weber et al. (2020) at UKHD to avoid the risk of incompleteness and lack of comprehensibility of relevant information (Weber et al., 2020) in contrast to FTORs, which are traditionally written in a continues text form. The implemented software application of Weber et al. (2020)

Patient Degree	SOR Category	German Template
complete response (CR)	no tumour burden evidence	Oncological regular findings without evidence of recrudesce or metastasis (Onkologisch regelrechter Befund ohne Nachweis von Rezidiv oder Metastasierung)
partial response (PR)	significant decrease of tumour burden	Oncological improvement of findings; constancy of findings with a tendency to decrease (Onkologisch Befundverbesserung; Befundkonstanz mit tendenzieller Abnahme)
stable disease (SD)	no significant change of tumour burden	Oncological constancy of findings (Onkologisch Befundkonstanz)
progressive disease (PD)	significant increase of tumour burden	Oncological worsening of findings; constancy of findings with a tendency to increase (Onkologisch Befundverschlechterung; Befundkonstanz mit tendenzieller Zunahme)

Table 3.1: *Example translation of the different TRCs to a SOR as executed by the software for structured oncology reporting. This table was adapted from Liang et al. (2022). SOR=Structured Oncology Report, TRC=tumor response category.*

provides a standardized, structured layout that comprises disease-specific report templates, a tabulated tumor burden documentation and a standardized conclusion. Figure 3.8 gives an example of a SOR. The conceptual design of the SOR templates follows a level 2 reporting structure (Fink et al., 2022a; Weber et al., 2020). i.e. the reports are created with a browser-based tool that provides drop-down menus and pick lists but also text forms that allow entering free-text information (Fink et al., 2022a) [1].

An important specification in oncological reports is the description of changes in tumor burdens during treatments. For this reason, a working group comprising multiple cancer care institutions published a set of rules called Response evaluation criteria in solid tumors (RECIST), which allow assessing the activity and efficacy of new cancer therapeutics in solid tumors using standardized terminology (Eisenhauer et al., 2009; RECIST, 2023). Following the RECIST version 1.1 guidelines all considered German reports can be classified into four tumor response category (TRC): PD, SD, PR, CR. The tool to create SORs by Weber et al. (2020) applies the standardized terminology RECIST version 1.1 guidelines and considers baseline and nadir imaging if applicable (Eisenhauer et al., 2009). The tool also allows to assign the TRC of "tendency of progressive disease (PD)" and "tendency of partial response (PR)". Since the TRC across reports are often not equally distributed and the complexity of reports of different TRC strongly varies, a major challenge for proposed algorithms is to handle the imbalances and heterogeneities in the datasets. Table 3.1 shows how the TRC are translated into the final finding by the software for structured oncology reporting.

The FTOR-DKFZ and FTORT-TKH follow a similar high-level structure consisting of a

---

[1] Online version of the software application can be assessed here: http://www.targetedreporting.com/-sor/, December 16th, 2022

Dataset	Institution	Case num	Field strength (T)	Resolution (in/ through plane)(mm)	Endorectal coil	Manufactor
Site A	RUNMC	30	3	0.6-0.625/3.64	Surface	Siemens
Site B	BMC	30	1.5	0.4/3	Endorectal	Philips
Site C	I2CVB	19	3	0.670.79/1.25	No	Siemens
Site D	UCL	13	1.5 and 3	0.3250.625/33.6	No	Siemens
Site E	BIDMC	12	3	0.25/2.23	Endorectal	GE
Site F	HK	12	1.5	0.625/3.6	Endorectal	Siemens

Table 3.2: *Site-specific technical parameters of the multi-site prostate MRI segmentation dataset. This table was adapted from Liu et al. (2020a). MRI=magnetic resonance imaging.*

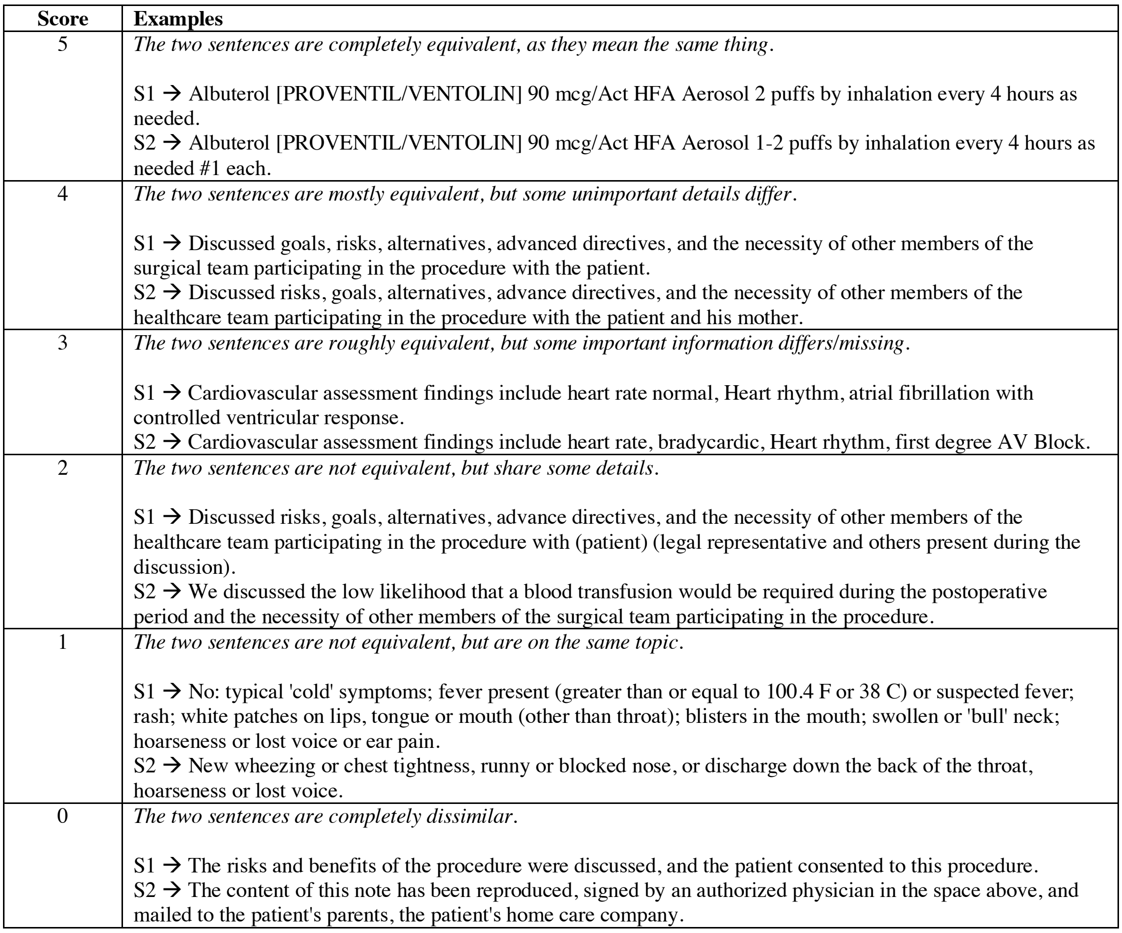
general information (except for the FTORT-TKH), imaging findings, and a conclusion (also denoted as "summary" or "impression") section. However, the FTORs are written more freely. The writing styles in reports vary strongly between clinical centers as shown in Section 5.3.1 in which the patient characteristics and the lexical complexity of the reports are analyzed in detail.

In principle, the general information section contains details on dates, the medical imaging device, information on the patient history and comparison to previous reports, and the general cancer treatment situation. The imaging findings section of an oncology radiology report contains information on primary tumors and metastases, their location and properties in different body regions, and other non-oncological findings such as fractures. The conclusion section concisely summarizes and assesses the patient's conditions based on the information given in the general information and imaging findings section.

Data curation, automated processing and manual annotation steps to use the radiology reports and to extract the TRC for the tasks of text classification and text summarization are given in Sections 4.3, 5.3.1 and 5.2.1.

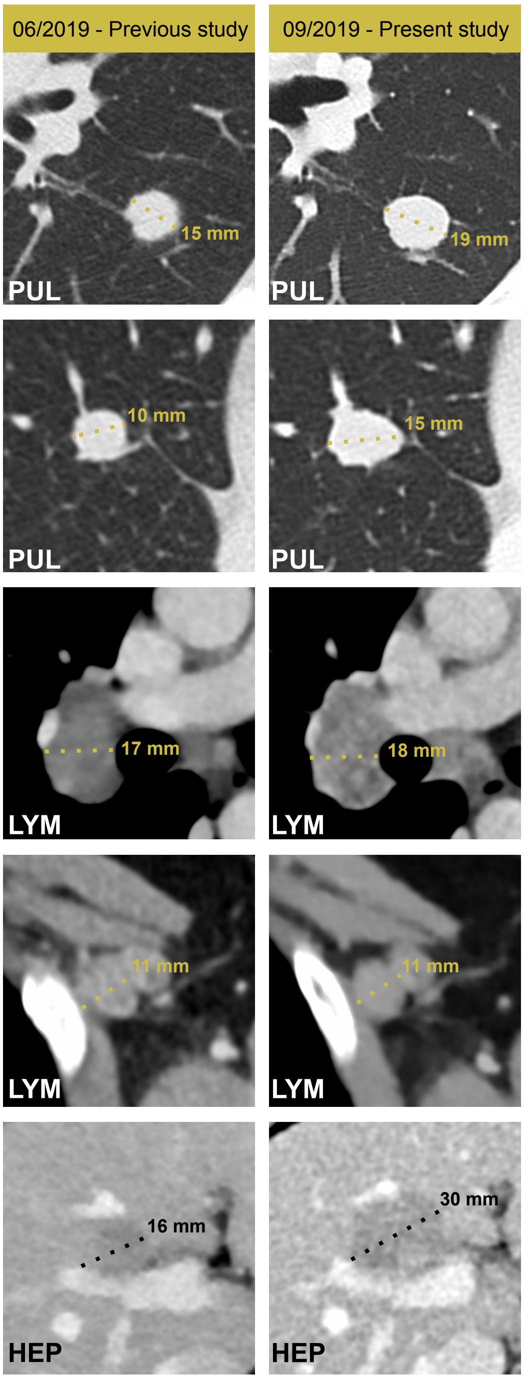
3.2.3 The multi-site prostate MRI segmentation dataset

The third dataset is an imaging dataset consisting of prostate T2-weighted MRI with respective reference segmentations of the prostate which include whole prostate segmentations as well as peripheral zone (PZ) and central gland (CG) segmentations. To evaluate domain generalizability and federated learning scenarios, the multi-site prostate MRI segmentation dataset is constructed based on six different data sources from previous challenges and benchmark datasets Liu et al. (2020a,b). The dataset contains samples of two different data sources from the NCI-ISBI 2013 Challenge (N et al., 2015; Clark et al., 2013; TCIA, 2023), one collection of samples from the benchmark I2CVB dataset (Lemaître et al., 2015; I2CVB, 2023) and samples of three different data sources from the PROMISE12 dataset (Litjens et al., 2014), which was created for the MICCAI Grand Challenge: "Prostate MR Image Segmentation 2012". Table 3.2 illustrates the details of the different data sources. In this work, an already pre-processed dataset (PROMISE12, 2023) is used, Liu et al. (2020a) resized each sample to 384x384 in the axial plane and normalized it to zero mean and unit variance. In addition, they clipped each sample to only preserve slices of the prostate region for consistent objective segmentation regions across all data sources (Liu et al., 2020a). Figure 3.9 illustrates samples of the preprocessed dataset for the respective sites.



Score	Examples
5	*The two sentences are completely equivalent, as they mean the same thing.* S1 → Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 hours as needed. S2 → Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 hours as needed #1 each.
4	*The two sentences are mostly equivalent, but some unimportant details differ.* S1 → Discussed goals, risks, alternatives, advanced directives, and the necessity of other members of the surgical team participating in the procedure with the patient. S2 → Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with the patient and his mother.
3	*The two sentences are roughly equivalent, but some important information differs/missing.* S1 → Cardiovascular assessment findings include heart rate normal, Heart rhythm, atrial fibrillation with controlled ventricular response. S2 → Cardiovascular assessment findings include heart rate, bradycardic, Heart rhythm, first degree AV Block.
2	*The two sentences are not equivalent, but share some details.* S1 → Discussed risks, goals, alternatives, advance directives, and the necessity of other members of the healthcare team participating in the procedure with (patient) (legal representative and others present during the discussion). S2 → We discussed the low likelihood that a blood transfusion would be required during the postoperative period and the necessity of other members of the surgical team participating in the procedure.
1	*The two sentences are not equivalent, but are on the same topic.* S1 → No: typical 'cold' symptoms; fever present (greater than or equal to 100.4 F or 38 C) or suspected fever; rash; white patches on lips, tongue or mouth (other than throat); blisters in the mouth; swollen or 'bull' neck; hoarseness or lost voice or ear pain. S2 → New wheezing or chest tightness, runny or blocked nose, or discharge down the back of the throat, hoarseness or lost voice.
0	*The two sentences are completely dissimilar.* S1 → The risks and benefits of the procedure were discussed, and the patient consented to this procedure. S2 → The content of this note has been reproduced, signed by an authorized physician in the space above, and mailed to the patient's parents, the patient's home care company.

Figure 3.7: *Examples of the sentence pairs from the BioCreative/OHNLP challenge 2018. This figure was adapted from Wang et al. (2018a). OHNLP=Open Health Natural Language Processing.*



GENERAL INFORMATION

Area: Chest (CT), Abdomen and Pelvis (CT)
Cancer treatment situation: Follow-up during treatment
Comparison: Last imaging study: 15.06.2019

ONCOLOGICAL FINDINGS

Primary tumor

Breast carcinoma on the left, primary diagnosis 2016, post breast-conserving surgery + ALNE + RTx, primary tumor region not sufficiently assessable on CT.

Metastases

Lung:

- Bipulmonary metastases slightly increased (ref. table).
- Unchanged pleural thickening left basal adjacent to the known nodule in the left lower lobe (2-145); pleural invasion cannot be excluded in this area.
- Status post VATS (S2 and S6 on the right side) with post-interventional scarring.

Thoracic lymph nodes and soft tissues:

- Mediastinal lymph node metastases slightly increased (ref. table).
- Constantly enlarged biaxillary lymph nodes.

Liver:

- Progressive and moderately contrast-enhancing lesion in liver segment SIVa (ref. table).

Abdominal lymph nodes and soft tissues: None.

Peritoneum: None.

Skeleton:

- Increasing demarcation of a mixed osteolytic-osteoplastic lesion in L1 (10-195) without posterior edge being involved.

Reference measurements (Follow-up):

1. 19 mm, prior 15 mm (PUL, left lower lobe, 2-80)
2. 15 mm, prior 10 mm (PUL, middle lobe, 2-123)
3. 18 mm, prior 17 mm (LYM, right hilar region, 5-82)
4. 11 mm, prior 11 mm (LYM, left axillar region, 5-49)
5. 30 mm, prior 16 mm (HEP, SIVa, 5-183) Sum of diameters: 93 mm, prior 69 mm (+35%)

NON-ONCOLOGICAL FINDINGS

Chest: Thyroid nodule in the right lobe measuring 6 mm (5-8).
Abdomen and Pelvis: Unremarkable.
Skeleton: Age-related degenerative changes.

IMPRESSION

Oncological impression

- Significant increase of tumor burden compared to 15.06.2019.
- Known bipulmonary metastasis with increasing lesions and one newly detectable lesion. No change in lymph node metastases.
- Solitary progressive liver metastasis.
- According to RECIST 1.1: PD (progressive disease).

Non-oncological impression

Constant thyroid nodule in the right lobe.

Figure 3.8: *Oncologic assessment in the clinical routine using structured oncology reporting. Shown is an exemplary SOR for a 32-year-old woman with a history of breast cancer (left side). The report is interpreted with the TRC PD. This figure was adapted from Fink et al. (2022a). PD=progressive disease, SOR=Structured Oncology Report, TRC=tumor response category.*

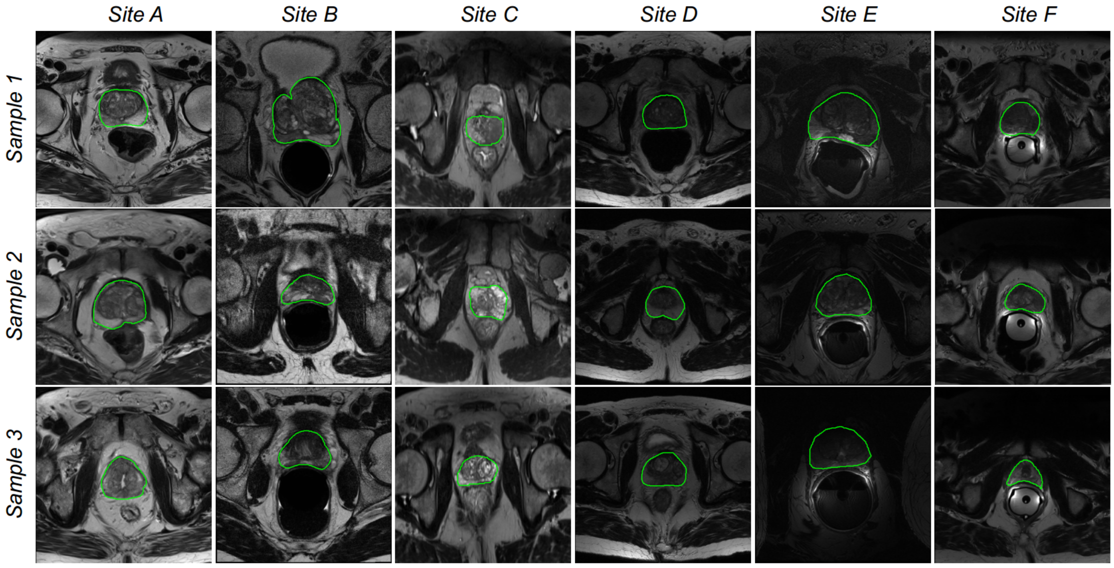


Figure 3.9: *Samples including the prostate ground annotation from the multi-site prostate MRI segmentation dataset. Taken from https://liuquande.github.io/SAML/, December 16th, 2022. MRI=magnetic resonance imaging.*

4 Methods

4.1 Semantic modeling for semantic textual similarity

In this section, the first part, Section 4.1.1, presents the baseline on the Clinical STS dataset, followed by an analysis of the dataset. The analysis motivates three approaches to improve the performance of the BERT baseline, which are presented in detail in the next part, Section 4.1.2. The metric to measure the performance of the different methods is the Pearson correlation coefficient (PCC), which measures the linear correlation between the predicted similarity scores and the annotated similarity scores (Kades et al., 2021).

4.1.1 Overview

Baseline

The baseline consists of the pre-trained ClinicalBERT model (Alsentzer et al., 2019) in combination with a linear regression layer. BERT was pre-trained using masked word prediction and next-sentence prediction. To facilitate the pretraining (Devlin et al., 2019) introduced two special tokens. A so-called Separator token (SEP token) was introduced to mark the end of an input sequence and a so-called Classifier token (CLS token) was introduced for the next sentence prediction task. Therefore, every input sequence of the BERT model starts with a CLS token and ends with a SEP token. Due to the n to n attention of the transformer architecture, the CLS token captures aggregated information of all tokens in the input sequence. For the task of STS an additional SEP token in combination with position encoding is used to distinguish the two input sequences. For the similarity regression task, the token vector representation of the CLS token in the last layer of BERT serves as input for an additional linear regression layer consisting of a single neuron. During training, the pre-trained BERT model and the linear layer are fine-tuned by minimizing the Mean Squared Error (MSE).

Cluster analysis

To determine weaknesses of the BERT baseline, a k-means clustering algorithm on the InferSent embeddings (Conneau et al., 2017) of the sentence pairs is performed. Subsequently, for each cluster, the absolute difference between the reference and the predicted scores from BERT are considered for each emerging cluster. The use of InferSent embeddings is motivated by its good semantic representation of sentences (Conneau et al., 2017). More details and the resulting visualizations are presented in the experiments and results in Sections 5.1.1 and 5.1.2.
The results suggest that for some sentence pairs, such as pairs belonging to cluster 3, the BERT model fails to predict the similarity score correctly. Samples from the cluster suggest that this concerns mainly sentence pairs that prescribe medication, for example, furosemide [LASIX] 40 mg tablet 1 tablet by mouth two times a day or "ondansetron [ZOFRAN] 4 mg tablet 1 tablet by mouth three times a day as needed. The analysis motivates the third approach (medication graph), which focuses solely on sentence pairs that prescribe medication.

4.1.2 Approaches

The three approaches to improve the BERT baseline are illustrated in Figure 4.1.

Enhancing BERT with features based on similarity measures

Before the introduction of transformer models, the task of STS was tackled using similarity measures applied on character and token level as well as to sentence and token embeddings. The pre-training of BERT is designed to create a model that contains a universal language understanding and semantic knowledge. The motivation of the first approach is combining different similarity measures, which might add additional information to the one BERT captures during pre-training and fine-tuning. Therefore, multiple token-based and sentence-embedding-based similarity measures are calculated. On a token level, n-grams of characters (a sequence of characters of length n) are created and compared using similarity measures like the Jaccard Similarity, which compares the proportion between the intersection and the union of n-grams in two input sequences. On a sentence level, the sentence representation between the two sentences is compared using similarity measures such as cosine

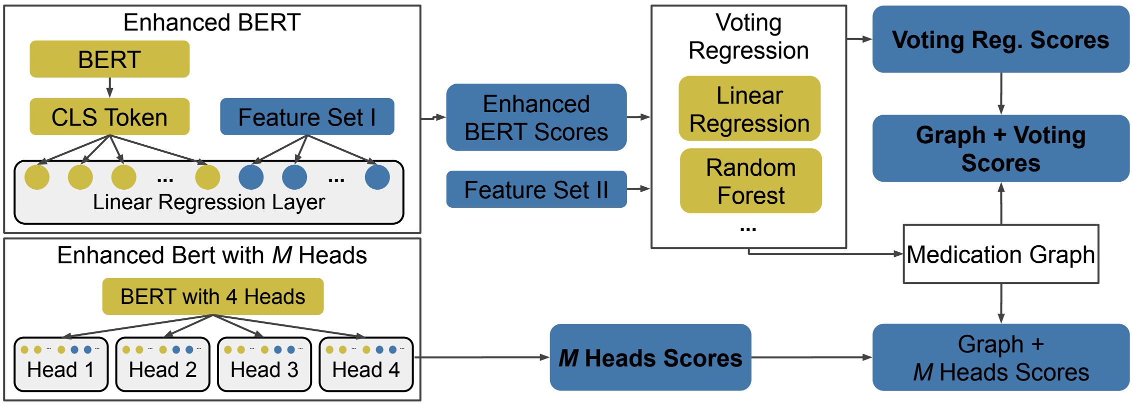


Figure 4.1: *Illustration of the processing pipelines for the different STS approaches. Calculated features such as Feature set I, Feature set II or just plain scores which serve as input for a processing step are colored blue. The framed boxes describe the four main processing steps. The scores marked in bold represent the scores that were submitted to the challenge. In the Enhanced BERT approach, the CLS token is highlighted which, in combination with Feature set I, serves as input for the linear regression layer in the STS downstream task. The medication graph takes only the subset of scores from sentences that prescribe medication as input. After its application, the subset of original input scores is replaced by the updated scores. This figure was adapted from Kades et al. (2021). BERT=Bidirectional Encoder Representation from Transformers, CLS token=Classifier token, STS=semantic textual similarity.*

similarity. The selection of the similarity measures builds upon the ones introduced by Chen et al. (2018).
The calculated features (Feature set 1 and Feature set 2) are added at two positions in the processing pipeline. In the first case (Enhanced BERT), the CLS token is extended by similarity measures from the Feature set 1, before feeding it to the linear regression layer. In the second case (Voting Regression), the similarity measures from the Feature set 2 are merged with the predicted output scores from Enhanced BERT and fed into a voting regressor (Pedregosa et al., 2018) consisting of several estimators. More details on the Feature sets and the estimators are given in Section 5.1.1.

M-Heads

The cluster analysis of the sentences shows the high variability in the dataset. Ensembling methods are common approaches for training a model to concentrate on different characteristics and properties of the dataset (Opitz and Maclin, 1999; Russakovsky et al., 2015). The main idea is to duplicate parts of the model or the whole model

and aggregate the predicted results. This way, different repetitions of the model can concentrate on other characteristics in the dataset. The final aggregation of the single predictions creates a group opinion over the predictions of the single models. Thereby, the aggregation avoids the dominance of a single model, mitigating the risk of just reacting to noise in the input data (Kades et al., 2021; Lee et al., 2015).
In the architecture of BERT, the most straightforward position to incorporate the ensembling methods is at the level of the linear regression layer. Therefore, the regression layer receiving the CLS token representation is duplicated multiple times, functioning as head of the ensembling approach. Before training, each head is initialized with different weights to avoid learning the same characteristics. Similar to Ilg et al. (2018) and Rupprecht et al. (2017) a loss scaling is employed to enforce specialization of the different heads (Kades et al., 2021).

Training:
During training, the loss scaling is implemented as follows. As described above, each head h with $h \in \{1, ..., M\}$ gets as input the CLS token token representation of the last layer of BERT. Each head h consists of a linear regression layer and outputs a similarity score s_h. The loss l_h for each head is calculated using the MSE between the predicted similarity score s_h and the reference similarity score $s*$:

$$l_h = MSE(s_h, s*) \tag{4.1}$$

The intuition behind the loss scaling is that the head with the lowest loss gets updated the most. The head with the lowest loss $h*$ is determined as follows:

$$h* = \underset{h=1,...,M}{\operatorname{argmin}}(l_h) \tag{4.2}$$

In the next step, the weight for each loss is determined with:

$$a_h = \begin{cases} \beta_1 & \text{if } h = h* \\ \frac{\beta_2}{M-1} & \text{if } h \neq h* \end{cases} \tag{4.3}$$

with $\beta_1 < \beta_2$. In this work, $\beta_1 = 0.95$ and $\beta_2 = 0.05$.
Then, the total loss $\mathcal{L}$ is calculated by a weighted mean of the losses from the heads:

$$\mathcal{L} = \frac{1}{M} \sum_{h=1}^{M} a_h \cdot l_h \tag{4.4}$$

Using the loss scaling, the head with the lowest loss gets updated most by a fraction of β_1. The other heads share the remaining fraction of β_2.

Prediction:
During inference, the predicted similarity scores s_h from each head are averaged to create a final similarity score s:

$$s = \frac{1}{M} \sum_{h=1}^{M} s_h \tag{4.5}$$

Medication graph

In this approach, only sentence pairs that prescribe medication are considered. They are denoted as "medication sentences" and include e.g. "ibuprofen 150 mg tablet 2 tablets by mouth every 7 hours as needed", with more examples in the discussion. The partially structured sentences allow for the analysis of individual entities, which comprise the active agent ("ibuprofen"), the strength ("150 mg"), does ("2 tables"), and frequency ("7"). The entities are automatically extracted using the MedEx-UIMA system (Jiang et al., 2014; Xu et al., 2010). From a medical point of view, the active agent significantly impacts the similarity between two sentences. For this reason, a way to model the similarity between active agents is presented in the following. In the second step, the calculated similarities between the agents are combined with the remaining entities.

The idea to model the similarity between active agents builds upon extrapolating the similarity information embedded in the training dataset. In detail, the similarity information of two active agents A and B, together with the similarity information of two active agents B and C, might provide knowledge about the similarity between agents A and C. By constructing a weighted graph with the active agents as nodes and the similarity scores as edges, similarity scores between two arbitrary active agents can be determined incorporating the similarity scores on the shortest path between the two considered agents. Before explaining the graph construction details, first, details of how the remaining entities are handled.

Feature construction:
The remaining entities are strength, dose, and frequency. Strength and dose can be further split up into ratio and nominally scaled entities, which results in a total of $N = 5$ entities. For example, the strength "4 mg" into the number "4" and the unit ("mg") and the dose "2 tables" into the number "2" and the unit "tables". Therefore, considering the entities of two sentence pairs $e_{k,1}$ and $e_{k,2}$, the similarity features Δ_k with $k \in \{1, ..., N\}$ can be calculated for nominally scaled entities such as units by applying comparison:

$$\Delta_k = \begin{cases} 0 & \text{if } e_{k,1} = e_{k,2} \\ 1 & \text{if } e_{k,1} \neq e_{k,2} \end{cases} \tag{4.6}$$

and for ratio scaled entities by calculating the squared difference:

$$\Delta_k = (e_{k,1} - e_{k,2})^2 \tag{4.7}$$

Graph construction:
For the graph construction, all medications sentences $S = (a_1, a_2, s, \Delta_1, ..., \Delta_N)$ form the training dataset are considered. Each sentence pair consists of the active agents a_1 and a_2, the similarity scores s, and the entity features Δ_k. A weighted similarity graph $G(V, E)$ is created by using all active agents A_i as nodes $V = \{A_1, A_2, ...\}$ and by adding edges $E = \{(A_i, A_j, w_{ij})\}$ between all nodes, A_i, that are connected with an edge weight w_{ij} to each other. The edge weight w_{ij} incorporates the similarity scores s and the entity features Δ_k as follows:

$$w_{ij} = \frac{1}{|C|} \sum_{S \in C} \left(s + \tanh\left(\sum_{k=1}^{N} \lambda_k \cdot \Delta_k + \lambda_0 \right) \right) \in [s_{min}; s_{max}] \tag{4.8}$$

The outer sum is necessary because multiple sentence pairs might share the same agents. Therefore, all sentence pairs S containing the same active agents form a set of sentence pairs C. λ_k weights the remaining features and is learned in a training process explained in the next section. λ_0 corresponds to a bias. The idea of the $\tanh(x)$ function is to limit the change of the similarity score s. The final weight w_{ij} is bounded by $s_{min} = 0$ and $s_{max} = 5$.

The motivation behind the formula for the edge weights is that the similarity score s is annotated based on the active agents and the remaining features Δ_k. However, the medication graph should only mirror the similarity between the active agents without the influence of the remaining features. Therefore, the idea behind the sum of the Δ_k weighted by λ_k is to alter the similarity score s so that w_{ij} models the actual similarity between the active agents.

Inference:
During inference, the aim is to calculate a similarity score based on the similarity between two active agents and the remaining features Δ_k.

The similarity between the two active agents is incorporated in the created medication graph. In the best case, a direct link exists in the medication graph for the active agents occurring in an unseen sentence pair. In this case, the actual similarity between the two agents is given by the calculated w_{ij}. However, since $G(V, E)$ is not a complete graph, many cases exist in which no direct link exists. Thanks to the medication graph, it is still possible to assume the similarity between two active agents, which did not occur together in the training set. In detail, the existing weights on the shortest path between two active agents A_1 and A_3, such as the weights between A_1 and A_2 as well as A_2 to A_3, can help to interpolate the edge weight between the target agents A_1 and A_3. In this work, the interpolation is done based on the formula for calculating the resistance of parallel circuits:

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + ... \tag{4.9}$$

with R_i being the individual resistances and R_{eq} the final resistance of a resistor connected in parallel. For the final resistance R_{eq} holds $R_{eq} \leqslant \min R_1, R_2, ...$ (Meschede, 2015). Therefore, the total resistance is always smaller than the individual resistance. This translates in our case to the fact, that an interpolated similarity score w^*_{ij} is always lower than any of the scores along the shortest path, which is the motivation to use this formula. In detail, if the similarity score on the shortest path is e.g. very low, then it should not be possible to give the interpolated similarity score w^*_{ij} a higher similarity score than the lowest on the path. This assumption is motivated by the fact that there is no knowledge in the medication graph that says anything about the increased similarity of the agents under consideration. The formula for an interpolated weight w^*_{ij} along a shortest path $(w_{i,p_{ij}(1)}, w_{p_{ij}(1),p_{ij}(2)}, ..., w_{p_{ij}(M),j})$ is given by:

$$\frac{1}{w^*_{ij}} = \frac{1}{w_{i,p_{ij}(1)}} + \frac{1}{w_{p_{ij}(1),p_{ij}(2)}} + ... + w_{p_{ij}(M),j} \tag{4.10}$$

with $p_{ij}(1)$, $p_{ij}(2)$, ..., $p_{ij}(M)$ denoting the indices of the M nodes on the shortest path between A_i and A_j.

To determine the final similarity score s_g between two arbitrary sentence pairs in the medication graph, the actual similarity score w^*_{ij} needs to be combined with the ones from the remaining features Δ_k. This is done with:

$$s_g = w^*_{ij} + \tanh\left(\sum_{k=1}^{N} \lambda_k \cdot \Delta_k + \lambda_0\right) \tag{4.11}$$

with λ_k and λ_0 the learned weights for the entity differences.

In Figure 4.2, an excerpt of the constructed graph is illustrated. The graph highlights the shortest path between the active agent's calcium and prednisone, along with its corresponding edge weights.

Learning λ_k:

The parameters λ_k determine the entity features' contribution to the active agents' similarity score, which form together the final similarity scores. The weights λ_k are

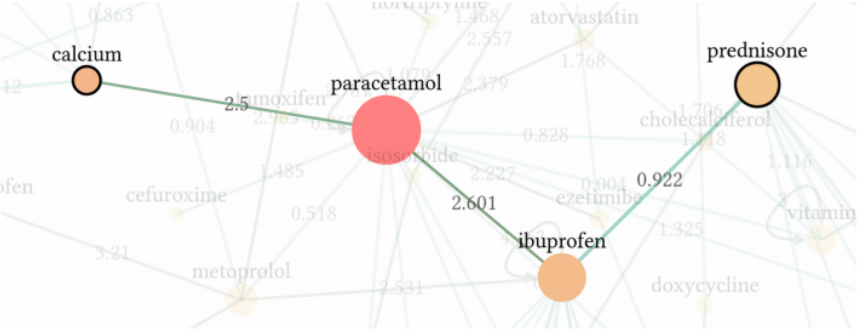


Figure 4.2: *Illustration of the medication graph, modeling similarities between agent pairs. In the excerpt, the shortest path with the corresponding similarities between two active agents is highlighted. The full version of the graph is available at https://med-graph.jansellner.net/, December 16th, 2022. In the online widget, further information about the calculations is provided and the similarity between two arbitrary agents can be visualized. This figure was adapted from Kades et al. (2021).*

learned using a random walk process by alternating the parameters λ_k until the graph performance is optimized. The graph performance is evaluated using a 10-cross-fold evaluation of the training data. During training, the MSE between the reference and the predicted scores is minimized. The MSE is preferred in this case over the PCC since a good PCC on a subset does not necessarily include a good performance on the complete dataset.

The mathematical formulation of the random walk is as follows. Let $\lambda = (\lambda_0, \lambda_1, ..., \lambda_N)$ be a vector of randomly initialized weights, with the aim to minimize $MSE(\lambda)$. During training, the weight of a randomly selected index is alternated using:

$$\lambda'_k = \lambda_k + N(0,1) \tag{4.12}$$

where N is sampled from a standard normal distribution. After one iteration, a new weight vector $\lambda = (\lambda_0, \lambda_1, ..., \lambda'_k, ..., \lambda_N)$ is created. The new weight vector is accepted if the following condition is fulfilled:

$$MSE(\lambda') < MSE(\lambda) \tag{4.13}$$

During training, random steps are repeated until there is no further improvement on the test folds. The learned λ_k parameters can then be used during inference.

Incorporating medication graph similarities into the two existing approaches:

Since the medication graph alone does not cover information about additional words or semantic relations, combining the scores with the two previous approaches is necessary. Using a Support Vector Regressor (SVR) the similarity scores of the medication graph s_g can be combined with the similarity scores from the previous approaches. While training the SVR, a Radial Basis Function is used as kernel and a regularization parameter C, as well as ϵ (ϵ-tube without penalty), are optimized.

4.2 Semi-structured data analysis for text summarization

This section adds two approaches to improve a transformer-based text summarization baseline. The first approach adds an extractive text summarization task to the model and the second approach leverages the pointer networks to copy words from the source to the target sequence. The presented models aim to automatically create the conclusion based on the radiology report's general information and imaging findings section (compare (cf.) Section 3.2.2). Zhang et al. (2018) emphasizes the importance of the general information section in the source text because it contains important information for the short-term and long-term examination of the patient's clinical record. For the text summarization task, the main challenge consists of transforming the salient and clinically significant sequences from a source of tokens $X = \{_1, x_2, ..., x_T\}$ to a concise sequence $Y = \{y_1, y_2, ..., y'_T\}$. All presented methods use the X and Y pairs of a radiology report for training a transformer-based $seq2seq$ model to generate Y (Liang et al., 2022).

The text summarization methods are evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) F1 score, the correct prediction of the TRC and by a human evaluation (Lin, 2004).

4.2.1 Approaches

Abstractive text summarization baseline

The BERT2BERT model proposed by Rothe et al. (2020) serves as the abstractive text summarization baseline. The model leverages the knowledge of pre-trained transformer models to reduce computational resources and to improve the sequence generation performance (Liang et al., 2022). Therefore, the encoder and decoder of the BERT2BERT model are initialized using a pre-trained BERT model. Only the encoder-decoder attention layers in the decoder need to be trained from scratch. Figure 4.3 illustrates the encoder-decoder structure of the BERT2BERT model. The baseline BERT2BERT model is trained using source and target sequence pairs (X,Y), with X representing the general information and imaging findings and Y the conclusion. For the input sequences X, the idea from Liu and Lapata (2019) to construct structured sequences is adopted. Therefore, the input X is constructed by concatenating the sentences from the general information and findings section separated by CLS tokens and terminated with a SEP token. The SEP token marks the end of the sequence so that the decoder stops generating tokens when it appears. Detailed information on the used model and the training parameters are given in Section 5.2.1.

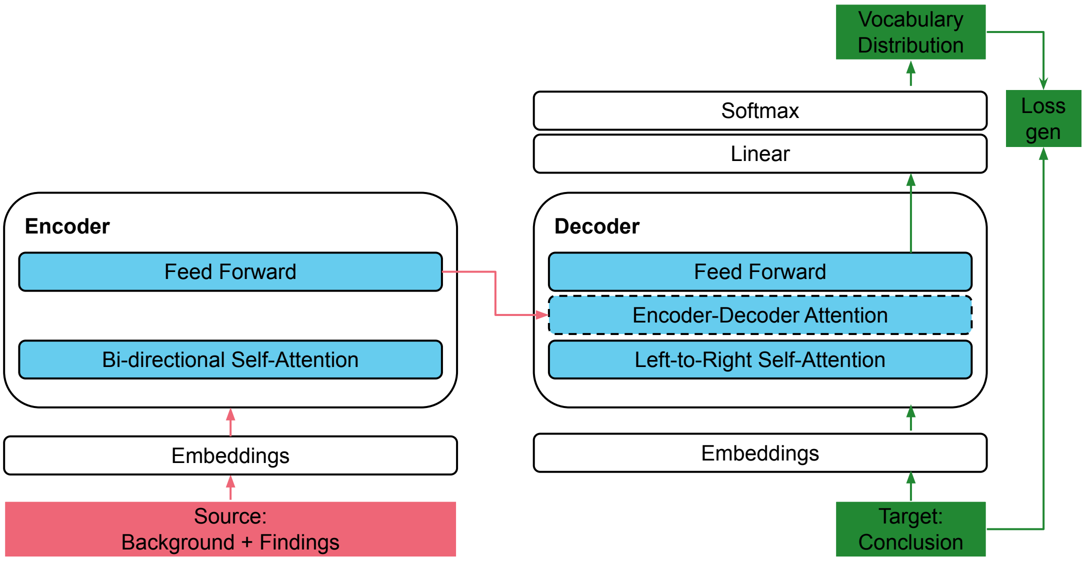


Figure 4.3: *Architecture of the BERT2BERT model. The encoder takes as input the general information and findings sections. The decoder generates the conclusion token by token. This figure was adapted from Liang (2021). BERT=Bidirectional Encoder Representation from Transformers.*

Baseline and extraction

Abstractive text summarization often struggles to create spurious facts, because of their ability to paraphrase (Liang et al., 2022). Combining extraction and abstraction approaches is one way to improve the correctness of the generated facts by directly extracting the facts from the source (Liang et al., 2022; Kryscinski et al., 2020; Cao et al., 2018; Zhang et al., 2020; Chawla et al., 2019; Falke et al., 2019). Commonly, the extraction and abstraction are incorporated separately into the model architecture (Hsu et al., 2018; Li et al., 2018a; Chen and Bansal, 2018). In this approach, instead of adapting the architecture, a new extraction learning task is added to the training of the BERT2BERT model. Therefore, the loss $\mathcal{L}$ for the extractive and abstractive optimization task is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{abstraction}} + \mathcal{L}_{\text{extraction}} \tag{4.14}$$

During training, the extractive task consists of generating key sentences from the source. Figure 4.4 illustrates the extension of the BERT2BERT with the extractive learning objective. The extractive and abstractive tasks are trained in parallel to both, training the model to reconstruct key phrases and generate new formulations from the source input. Therefore, the decoder receives, besides the conclusion Y, the key

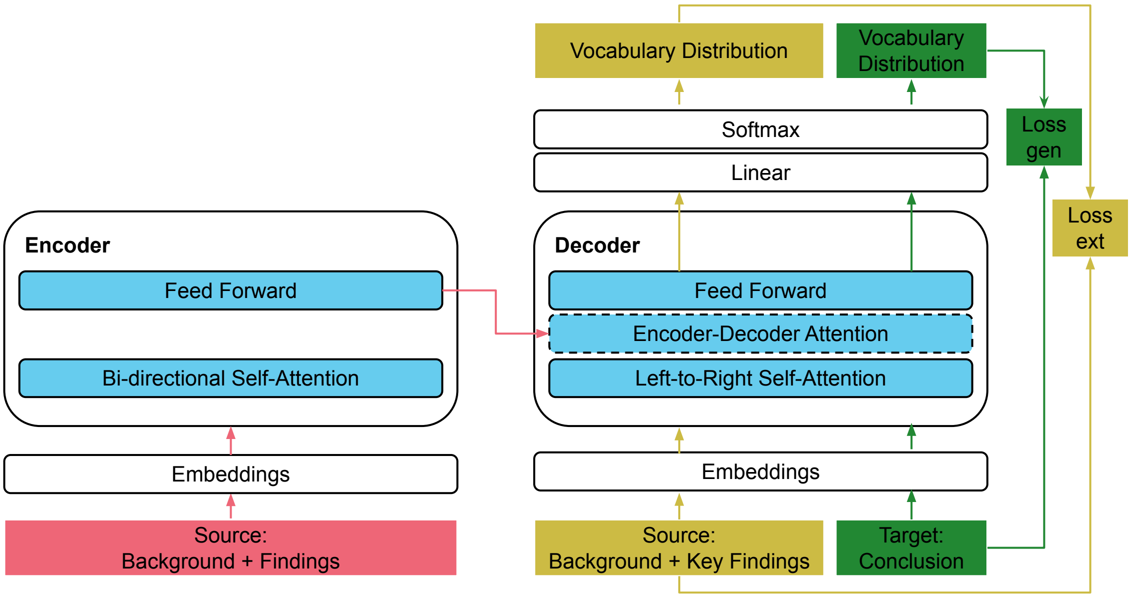


Figure 4.4: *Architecture of the BERT2BERT model with the extraction loss. The decoder is trained to generate besides the conclusion also the general information and key findings of the source input. During training, the two objectives are optimized simultaneously. This figure was adapted from Liang et al. (2022). BERT=Bidirectional Encoder Representation from Transformers.*

sentences X_{key} from the findings section as input.

Extracting key sentences

The advantage of the introduced extractive learning objectives is that no additional annotation is needed. However, the question is how to identify key sentences from the source. For this, three non-neural network-based but automated methods are evaluated.

- **Longest-**k: In this approach, the longest k sentences in the input serve as key sentences. The idea behind Longest-k is that the longer the sentence in the findings is, the more information it may contain for the summary.
- **TF-IDF-Ex**: In this statistical approach, the sentences are selected based on a TF-IDF analysis. A TF-IDF analysis (SPARCK JONES, 1972)relates the frequency of a term t in a document with the inverse document frequency (the number of documents that contain the term t) (see also Section 4.3.3). The resulting analysis outputs a weight for each term, describing its relevance. The analysis is applied to all radiology reports in the training datasets. The sentences are then

ranked by the highest sum of TF-IDF-weights per sentence. The sentences with the highest rank are considered key sentences.

- **TextRank** Mihalcea and Tarau (2004): In the TextRank algorithm, sentences are scored based on graph theory. Each sentence in the documents serves as the vertex of a constructed graph. The weight of the edges between two sentences is determined by their similarity, i.e. the number of overlapping tokens.

Therefore, all presented key sentence extraction approaches rank the sentences within the findings section according to their respective importance.

Baseline and pointer

The use of a pointer network is inspired by copying words such as dates or numbers, directly from the source to the target sequence (See et al., 2017). It comes with the benefits of increasing the factual correctness of the generated phrases and as an alternative generation method in case the BERT model is insecure in its generation. Pointer networks are often used in conjunction with RNNs, the usage in combination with a transformer encode-decoder model for summarization tasks is still the subject of research. The BERT2BERT model, together with the pointer mechanism, is illustrated in Figure 4.5. The original pointer network as presented by (See et al., 2017) consists of a linear layer and a sigmoid function and is defined as follows:

$$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_y^T y_t + b_{ptr}) \tag{4.15}$$

where vectors w_{h^*},w_s,w_y and the scalar b_{ptr} are learnable parameters and σ the sigmoid function (See et al., 2017). The network outputs the generation probability, p_{gen}, stating whether the next token is generated from the vocabulary distribution or the source input. In the formula, h_t^* is a so-called "context vector", which is created by a weighted sum of the encoder hidden states, where the weights are given by an attention distribution. s_t is the "decoder state", which is the output state of the decoder and y_t, the "decoder input", which corresponds to the word embedding of the current decoder input at time step t. (See et al., 2017) used a single-layer unidirectional LSTM as decoder.

To use the pointer network with the BERT2BERT baseline a few adjustments are necessary. For the creation of the context vector, an attention distribution is needed. In (See et al., 2017), the attention scores are given by the encoder-decoder attention layer. In the BERT2BERT model, multiple heads lead to multiple attention distributions, which can be combined into one attention distribution by taking the mean over the attentions of the heads (Deaton, 2019). Analogous to See et al. (2017), the context vector h_t^* at decoding step t can then be defined by weighting the encoder hidden states

h_j^* of the last encoder layer with the attention distribution a_t from the encoder-decoder states:

$$h_t^* = \sum_{j}^{J} \sum_{i}^{N_{heads}} a_t \cdot h_j^*, \tag{4.16}$$

with i the index of the attention head, j the position of the source sequence and J the total length of the source sequence (Liang et al., 2022). Like in (See et al., 2017), in the BERT2BERT the embeddings of the decoder input y_t as well as the decoder hidden states, s_t, of the last decoder layer serve as input for the pointer network.
The final probably distribution $P_{final}(w)$ to determine the next generated word w can be defined using the generation probability p_{gen} in combination with the vocabulary distribution $P_{vocab}(w)$ and the attention distribution as follows:

$$P_{final}(w) = p_{gen} \cdot P_{vocab}(w) + (1 - p_{gen}) \cdot \sum_{i:w_i=w} a_i^t \tag{4.17}$$

The dimension of $P_{vocab}(w)$ is the same as the vocabulary used by BERT. The attention distribution a_t has the dimension of the source sequences with corresponding index i in the vocabulary dimension. Since the encoder and decoder of BERT2BERT use the same vocabulary, the contributions from $P_{vocab}(w)$ and a_t can be summed up at the same indices.

Baseline combining extraction and pointer

The last model combines the extraction and pointer network approach.

Greedy search, N-gram blocking and repetition penalty

In all experiments, a greedy search is used in the decoding process to select the next token. Greedy search chooses the token with the highest probability at each step during the generation of the sequence. Preliminary experiments showed that greedy search is preferred over Beam search. More details are given in (Liang, 2021). Furthermore, to reduce repetition during the generation process, N-gram blocking and a repetition penalty are introduced, however, only during inference. N-gram blocking avoids the creation of a repetitive sequence of length N by identifying and replacing the repetitive token with a better candidate. The repetition penalty proposed by Keskar et al. (2019), prevents repetition by introducing a probability distribution for predicting the next token. The probability distribution takes into account the sequence of already generated tokens, based on which repetitive tokens receive a lower probability to be generated again. Detailed information on the N-gram blocking and a repetition penalty is given in (Liang, 2021).

4.2.2 Human evaluation

Text summarization methods are commonly evaluated using the ROUGE F1 score, which measures the overlap between a generated and reference summary. Since the ROUGE metric only covers a statistical-based evaluation of the text summarization methods, similar to Zhang et al. (2020), Sotudeh Gharebagh et al. (2020) and Nguyen et al. (2020), an expert evaluation with human annotators is conducted with the main aim to assess the clinical validity and the comprehensibility of the generated conclusion. For oncological reports, the clinical validity can be subdivided into oncological and non-oncological findings that are included in the conclusion. Besides the clinical validity, the expert annotators are asked to evaluate the comprehensibility of the generated summaries with a focus on the correct usage of medical terms and the readability of the summaries. This is especially interesting because summaries often contain acronyms, nonstandard clinical jargon and grammatically not complete sentences (Hasan and Farri, 2019). For a more fine-grained analysis, the three main criteria, oncological and non-oncological correctness, and comprehensibility, are evaluated in dependency of the tumor response categorys (TRCs): progressive disease (PD), stable disease (SD), partial response (PR), complete response (CR), which were introduced in Section 3.2.2.
The human evaluation is only applied to a subset of the test data. For a fair comparison of the four models, in the first step, a pool of samples from the entire test set is created for which all methods score higher than the overall average ROUGE F1 score. In the second step, five samples per TRC are randomly selected from the pool of samples, which creates a subset of 20 evaluation samples.
The evaluation is done with the annotation tool Doccano (Nakayama et al., 2018), which is hosted within Kaapana. The annotation was executed by one radiologist (M.A.F. [in training]) and one senior medical student (P.F. 12th semester). For the evaluation, the annotator is presented with the general information and imaging finding sections along with five random ordered summaries, with four system-generated conclusions and one reference conclusion. The motivation for the random order is, that the annotators have to compare the summaries relatively amongst each other instead of only to the reference. The annotators were asked to rate each summary with a Likert-scaled score from 0 (very poor) to 5 (very good) for the three criteria oncological and non-oncological correctness, and comprehensibility. The choice of a Likert-scaled approach instead of a rank-based approach was chosen because too often the generated conclusions ranked the same and a rank-based score can still be calculated using the Likert-scaled scores. The annotation guide for the annotators is given in the appendix in section B.1.2.

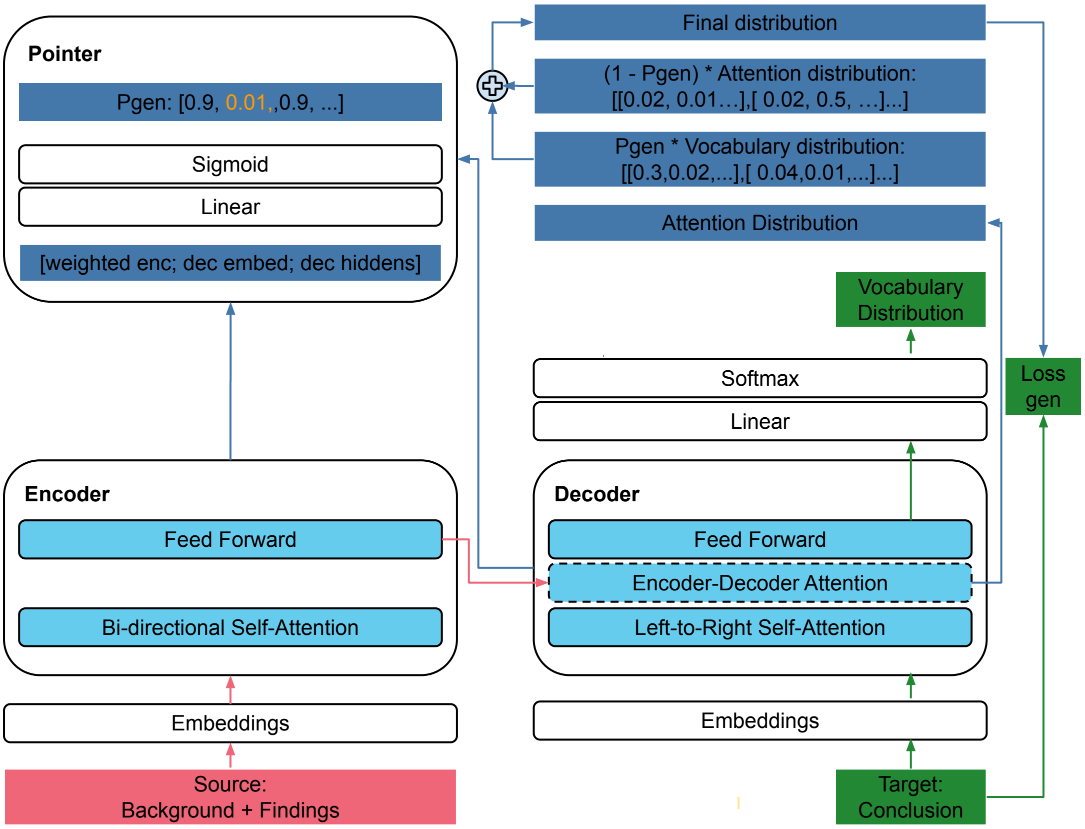

Pointer
Pgen: [0.9, 0.01,,0.9, ...]
Sigmoid
Linear
[weighted enc; dec embed; dec hiddens]
Final distribution
(1 - Pgen) * Attention distribution: [[0.02, 0.01…],[0.02, 0.5, …]...]
Pgen * Vocabulary distribution: [[0.3,0.02,...],[0.04,0.01,...]...]
Attention Distribution
Vocabulary Distribution
Softmax
Linear
Loss gen
Encoder
Feed Forward
Bi-directional Self-Attention
Embeddings
Source: Background + Findings
Decoder
Feed Forward
Encoder-Decoder Attention
Left-to-Right Self-Attention
Embeddings
Target: Conclusion

Figure 4.5: *Architecture of the BERT2BERT model incorporating the pointer network. The pointer network takes as input the weighted encoder hidden representation, the word embeddings of the decoder input token and the decoder hidden states at decoding step* t*. The pointer network generates a probability* p_{gen}*, which determines whether the next token is generated from the vocabulary distribution or the source input. This figure was adapted from Liang et al. (2022). BERT=Bidirectional Encoder Representation from Transformers.*

4.3 Assessing distributional shifts for text classification

This section aims to assess the performance of classifying the TRC for a dataset consisting of independent radiology reports from three different clinical sites. As presented in Section 3.2.2 the three datasets are different in structure, writing style and length. Therefore, the proposed text classification methods are mainly evaluated in terms of their generalizability and robustness, i.e. their ability to handle the distributional shift between the datasets. The performance of the automated text classification algorithms is evaluated against human baselines from annotators of different clinical expertise. In addition, it is evaluated how combining different generated annotations, such as the ones from a human and a NLP model, could improve the overall annotation performance. Furthermore, a section is dedicated to the explainability and interpretability of the applied AI methods, due to the high importance of understanding the medical decision processes. Before the application of the text classification algorithms, the datasets need to be prepared and annotated. This includes the curation of the data, a complexity analysis and the creation of a reference annotation. Information on the datasets is given in Section 3.2.2, and details on the data selection and curation are given in Section 5.3.1.

The human reference and baseline annotations are compared using the inter-rater readability metric Cohen's kappa (cf. Section 5.3.1). All classification performances of the methods are measured using the weighted F1.

4.3.1 Analysis of report corpora

Medical characteristics of the report corpora

While collecting the reports from the RIS, also metadata such as age, sex and acquisition date are retrieved. To additionally get an overview of the differences in the used data, the tumor families are retrieved. To make this as effective as possible a list of keywords describing different tumor categories is manually created, while the tumor categories are assigned to tumor families. Applying a regex with the keywords of the list to all documents gives a rough overview of the tumors occurring in the considered reports.

Complexity analysis of report corpora

A lexical complexity analysis is applied to the data to describe the distributional shift numerically and to evaluate the human and automated machine performance dependent on the different complexity characteristics. The lexical complexity analysis on the text classification input follows the extracted features describe in (Zech et al., 2018). Calculated complexity features include the word count, number of unique words, number of unique bigrams, type-token ratio, Yule's I metric and a so-called BERT split factor (Zech et al., 2018; YULE, 1939; Oakes, 1998). The type-token ratio of a text sample is defined as the ratio of the number of unique words V over the total token count T:

$$\text{type-token ratio} = \frac{V}{T} \tag{4.18}$$

Following Oakes (1998), the Yule's I metric is defined in this work as:

$$\text{Yule's I} = \frac{T * T}{\sum_i^V (c_i * c_i) - T} \tag{4.19}$$

with T the total token count and V the number of unique tokens i in the considered text sample. c_i denotes the frequency of the token i in the considered text sample. It measures the rate of how often words are repeated in a text. The higher the rate, the more complex is a text corpus.

The BERT split factor is introduced in this work to evaluate, whether the performance of the BERT-based text classification algorithm is influenced if words are not part of the vocabulary of BERT. The pre-trained BERT has a fixed vocabulary dictionary. In case the BERT tokenizer is not familiar with the word, it is split until all split units are part of BERT's vocabulary dictionary. Since German radiology reports contain a lot of words that are not part of the vocabulary dictionary, it is often necessary to split a word into multiple units. The BERT split factor is defined as the ratio between the number of tokens before and after applying the BERT tokenizer to a sample.

4.3.2 Reference annotations

As described in the dataset Section 3.2.2, a common terminology to describe the change in tumor burden are TRC. Due to the structure incorporated in the SORs, it is possible to automatically extract the TRC from most of the reports. However, for the FTOR-DKFZ and FTORT-TKH a manual reference annotation is necessary.

Automated reference annotation from SORs

As described above the reporting system from Weber et al. (2020) follows the RECIST guidelines to describe changes in the tumor burden. Therefore, the TRC can be automatically retrieved using a regular expression. The tumor response categories containing tendencies are converted to the four TRC by removing the tendency description.

Manuel reference annotation of FTORs

For the FTORs a manual reference annotation is necessary. For this, the two radiologists (M.A.F. [in training] and J.K [board certified]), with five and six years of experience in oncologic imaging independently reviewed all retrieved FTORs in random order (Fink et al., 2022a). The labels were created using again the text annotation tool Doccano from Nakayama et al. (2018) hosted via Kaapana. The annotators were presented with the respective FTORs from the DKFZ and the TKH in random order. The TRC can be determined based on the "findings" and "impression" sections of the report. Additionally, for the FTORT-TKH, non-oncological labels (worsening, constant, improving) were created, describing the change in the non-oncological findings (e.g. "increase in degenerative changes of the spine"). A consensus review round after the first annotation round resolved occurring disagreements between the two radiologists (Fink et al., 2022a).

4.3.3 Text classification baselines

The created reference annotations on the FTORs are used to assess the performance of classifying the TRC solely based on the "findings section" of a report. The performance of human annotators with different medical expertise is compared against statistical and transformer-based text classification algorithms. In addition, the performance of ensembles of different human and machine-based methods are compared.

Human baseline on FTORs

Similar to the creation of the reference annotations, the annotation tool Doccano hosted by Kaapana was used for the human annotations. This time the annotators were only presented with the findings section, excluding the general information and the impression sections. In total, the dataset was annotated by seven annotators: two radiologists (A.B [in training] and M.M [board certified], with 4 and 6 years of oncological imaging experience), two medical students (M.S. and M.K., third and 12th semesters, respectively), and three radiology technologists (RTs) students (all third semesters), were asked to assign a TRC to the randomly ordered FTORs. Furthermore, the annotators were asked to assign a confidence score to the selected TRC. The confidence was scored using a five-point Likert scale (1 = not confident at all; 5 = very confident).

Text classification algorithm

The text classification algorithms are developed based on the SORs. This way, no manual annotation is needed, since the reference annotations for the SORs are automatically retrieved. Two types of NLP models are presented to assess how they handle the distributional shift to the FTORs. The models take as input for training and testing only the findings section of a report excluding the general information and impression section.

BERT for text classification

The first algorithm builds on top of a German pre-trained BERT model (Devlin et al., 2019; deepset, 2023). In the pre-training step, BERT integrates the CLS token for the next sentence prediction task. This token can be exploited to be used for the task of text classification. The token is suitable because it is present in every sequence by design and through the n to n attention, it contains the semantic meaning of the whole sequence. For the task of text classification, the CLS token representation of the last layer of BERT is fed into a linear classification layer. During fine-tuning, the model is trained using a cross-entropy loss. Since the data in this work have imbalanced class distributions a weighted cross-entropy loss is applied:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(p(x)_c) \tag{4.20}$$

with C the number of classes, y the reference class, x the output scores of the linear classification layer, w the class weights, and p the softmax function. p is defined as follows:

$$p(x)_c = \frac{e^{x_c}}{\sum_{j=1}^{C} e^{x_j}} \tag{4.21}$$

One main drawback of the BERT model is its memory consumption of $\mathcal{O}(n^2)$ with n denoting the maximal number of tokens for the input sequence. For the pre-trained BERT used in this work, the maximum sequence length is 512 tokens. It must be noted, that due to the unknown medical vocabulary, many words need to be split by the BERT tokenizer, which might transform actual short sequences into long sequences. As discussed in the related work Section 2.3, many approaches exist to handle sequences longer than 512. Since most of the newer models are only available in English, this work uses the approach proposed by (Mulyar et al., 2019), who splits long sequences into chunks of 512 tokens. A vector representation of the whole input sample, which serves as input for the linear classification layer, is then given by averaging the CLS token representations of the chunks. However, unlike (Mulyar et al., 2019), only the first chunk for backpropagation is used in the training phase with the purpose to keep the GPU memory footprint constant, independent of the input sequence length. The setup is illustrated in Section 2.3.

Conventional NLP models

Besides the state-of-the-art model, three feature-rich NLP models are evaluated. The approaches consist of a TF-IDF analysis followed by different classification models. The TF-IDF (SPARCK JONES, 1972) for a document d and a term t is defined as follows:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t) \tag{4.22}$$

with $\text{tf}(t, d)$ denoting the raw count of a term t in a document d and the inverse document frequency:

$$\text{idf}(t) = \log \frac{1 + n}{1 + \text{df}(t)} + 1 \tag{4.23}$$

where n is the the total number of documents and $\text{df}(t)$ the number of documents that contain the term t. The output of the analysis is a feature vector containing the TF-IDF weights for a fixed vocabulary. In this work, the vocabulary is defined by the tokens with the highest term frequency, tf, across the corpus. The generated feature vectors are further processed by a classification model. Following Das and Chakraborty (2018), a Linear Support Vector Classifier (Linear-SVC), K-nearest neighbors (KNN) and Multinomial Naïve Bayes (MNB) are selected for the text classification task.

For the training of all models, a five-fold ($k = 5$) cross-validation was performed along with an extensive hyperparameter search. Details on the experiments along with hyperparameter settings for the TF-IDF analysis and the different classification models are given in Section 5.3.1.

4.3.4 Interpretability and explainability of the models

Especially in the medical domain, the interpretability and explainability of machine learning models are of high importance. To qualitatively assess the decision process of the proposed methods, tokens responsible for a model's decision can be highlighted in the source text (Arras et al., 2017; Wiegreffe and Pinter, 2019; Huang et al., 2020). Furthermore, the feature vectors which are fed to the classification algorithms are plotted using dimension reduction methods. Finally, for all algorithms, methods are implemented to provide the model's confidence for a selected class.

Highlighting important tokens

BERT's attention

As elaborated in the related work Section 2.3, the interpretability of BERT's attention is the subject of research. In this work, the attention weights of the tokens before and after fine-tuning the BERT model are visualized. To make the subset of attention weights comparable across samples of different sequence lengths, each attention weight w_i is rescaled by the ratio between the sequence length N and the maximum allowed sequence length $N_{max} = 512$ of the BERT model:

$$w_i^* = w_i * \frac{N}{N_{max}}.$$

Otherwise, short sequences would have higher attention weights compared to long sequences. In the following, the scaled attention weights w_i^* are denoted as "token importance". For words that were split by the BERT tokenizer, the token weights are averaged and put again together to the original word.

TF-IDF weights

For the TF-IDF-based approaches, the contribution of tokens to the decision can be visualized using the TF-IDF weights (Arras et al., 2017).

Vizualization of the feature vectors

This section tries to qualitatively assess the knowledge about the different classes incorporated into the feature vectors that are fed into the classification models. For the fine-tune BERT model this corresponds to the averaged CLS token embeddings, representing a sequence. For the TF-IDF-based algorithms, this corresponds to a vector of TF-IDF weights. The feature vectors are visualized using the dimension reduction techniques Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (Hinton and Roweis, 2002; van der Maaten and Hinton, 2008; McInnes et al., 2018).

Calibration of the model's predictions

Especially in medical decision processes, it is not only important to obtain accurate predictions, but also indications of how likely a decision is incorrect (Guo et al., 2017). Therefore, the probabilities for different classes provided by the models should reflect the likelihood of a correct prediction. In detail, the probabilities in a multi-class classification problem correspond to calibrated confidences if the given probability assigned to a class matches the true number of cases in which the prediction was correct. For example, given 100 predictions with each one predicting class A with a probability of 0.8, then the true number of correct cases should be 80 (Guo et al., 2017).

To which extent probabilities generated by machine learning models are calibrated confidences can be calculated using the so-called ECE (Guo et al., 2017). Considering a set B_m of indices of samples whose prediction confidence falls into the interval

$I_m = \left[\frac{m-1}{M}, \frac{m}{M}\right]$ of M equally spaced bins, then the accuracy of B_m is given by:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i), \tag{4.24}$$

with $\hat{y}_i$ and y_i the predicted and true class labels for sample i. The confidence of B_m is then given by:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \tag{4.25}$$

with $\hat{p}_i$ the confidence for sample i. Following Guo et al. (2017), the expected calibration error is then defined as follows:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \Big|, \tag{4.26}$$

with n the number of samples.

The general idea is to group the predictions into bins. Then, the differences between the accuracy of the predictions and the averaged probabilities per bin are summed up. Therefore, the lower the ECE the better the probabilities are calibrated.

The BERT-based approach outputs probabilities after applying a softmax to the logits generated by the linear classification layer. To transfer the probabilities to calibrated confidences, a so-called temperature scaling presented by Guo et al. (2017) is applied to the generated probabilities. (Desai and Durrett, 2020) has proven the successful application of temperature scaling on multiple downstream tasks. To ensure that For the probabilities generated by the classical classification algorithms, Linear-SVC, KNN and MNB, generation of calibrated probabilities is enforced using the "CalibratedClassifierCV" (scikit-learn, 2023) from scikit-learn (Pedregosa et al., 2011), which ensures the creation of calibrated confidences using a cross-validation approach.

4.4 Manual data annotation and tagging for medical images

The performance and generalizability of AI algorithms heavily depend on the quality and the amount of the training data (Huynh et al., 2020). Therefore, a major challenge consists in creating sufficiently large, curated, and representative training data (Willemink et al., 2020). The creation of annotated, well-curated data requires experts and is a time-intensive process. This section presents a software solution for efficient DICOM-image tagging and cohort curation (Kades et al., 2022a). The presented implementation is evaluated on the feasibility of multiple use cases which are outlined in Section 5.4.1. The use cases include the creation of a well-curated and tag-based annotated dataset.

4.4.1 The used open-source software toolkits

The toolkit is created leveraging multiple open-source tools, which are all integrated into the medical imaging provisioning platform toolkit Kaapana, which is presented in Section 3.1.1 in detail. Kaapana uses a Kubernetes cluster to host micro-services and to spawn up on-demand processing containers. The main contribution in this section consists in optimally using and combing tools integrated into Kaapana for the tasks of cohort curation and image tagging. The tools used include the following:

Internal PACS
The platform integrates an open source PACS system dcm4chee, which allows using a DICOM receiver port to directly receive images from a clinical PACS. The platform internal PACS serves as the main storage for DICOM objects within a Kaapana platform.

Elasticsearch and Kibana
The tools Elasticsearch and Kibana are used for managing the metadata of the stored DICOM objects. Upon the arrival of DICOM images, their metadata is extracted and stored in the search engine Elasticsearch. The dashboard solution Kibana allows the viewing and filtering of the retrieved metadata. The provided filtering option allows the creation of a cohort for potential AI applications solely based on the metadata of the DICOM objects. The dashboard of Kibana for filtering the DICOM metadata is denoted as the "meta dashboard".

Airflow
The pipelining tool Airflow is integrated as a central workflow management system. It allows the creation of customized processing pipelines that are applied to the medical

image data. The integration with the Kubernetes cluster allows running on-demand processing containers (Scherer et al., 2020; Kades et al., 2022a).

OHIF Medical Imaging Viewer:
The zero-footprint web-based OHIF Medical Imaging Viewer (Ziegler et al., 2020) allows displaying DICOM objects such as images and corresponding pixel-wise annotations. DICOM objects stored in the internal PACS are directly accessed using the DICOMweb standard (Ziegler et al., 2020). The tool also supports various image manipulation and annotation tasks.

Doccano
The open-source text annotation tool Doccano (Nakayama et al., 2018) offers various annotations features such as text classification, sequence labeling, and sequence-to-sequence tasks (Nakayama et al., 2018). The tool is implemented using a Django backend (Django, 2023) with a PostgreSQL database (PostgreSQL, 2023) and a Vue.js frontend (Vue.js, 2023), which allows easy customizations. In its setup, the web-based tool uses projects for different annotation tasks. It supports a separation of users and provides an overview of the overall annotation process. Text data in various formats can be uploaded and annotated. For text classification tasks, custom color-coded labels can be created. Keyboard shortcuts for the labels allow an efficient annotation of the text data. Besides text annotations, Doccano also supports speech-to-text and image classification annotations. However, the annotation of medical images is hampered since Doccano only supports the use of images in the Portable Network Graphics format, which is seldom used for medical images and is limited to 2D images only. The annotation of medical images requires the support of DICOM objects and the provisioning of medical-accustomed navigation and interaction tools.

4.4.2 Combining the different toolkits

Integration of the OHIF viewer into Doccano
A central building block for the curation and tag-based annotation of medical images is the extension of Doccano with the OHIF viewer to display DICOM objects. Since the image view in the OHIF viewer is based on the Study Instance UID DICOM attribute (0020,00D), the proposed solution is limited to an annotation on the study level. For this, the source code of Doccano's user interface (UI) for annotations is modified to display an inline frame (iframe) containing the OHIF viewer instead of the textual input. In detail, the HyperText Markup Language (HTML) attribute, responsible to render the source text, is adapted to trigger the integration of the OHIF viewer, if it recognizes a specially formatted text containing the Study Instance UID DICOM attribute (0020,00D). The proposed substitution of medical images into

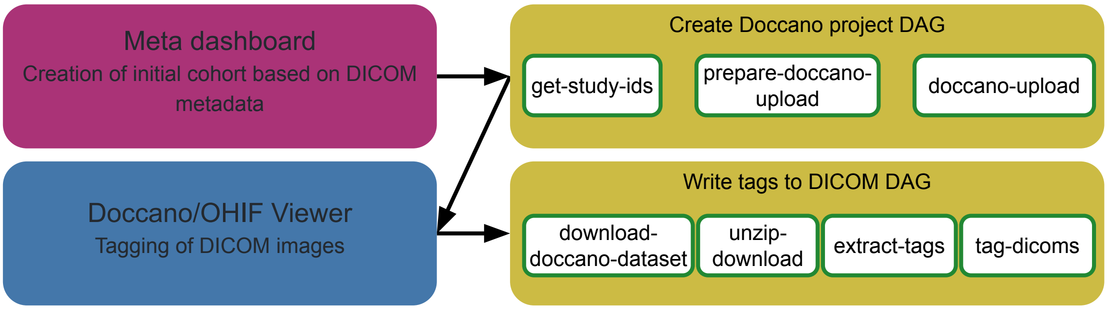


Figure 4.6: *Illustration of the annotation process within the Kaapana platform. The meta dashboard allows the creation of a cohort based on the metadata of the DICOM image. Airflow is used to process the cohort and to create a dedicated project on Doccano in which, with the help of the OHIF viewer, the DICOM images can be annotated. Once the annotation process is finished, another Airflow pipeline is triggered to persistently save the generated information on the platform. This figure was adapted from Kades et al. (2022a) with permission of Springer. DAG=Directed Acyclic Graph, DICOM=Digital Imaging and Communications in Medicine, Kaapana=Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis, OHIF=Open Health Imaging Foundation.*

the text attribute allows tagging and adding free-text data referenced by the Study Instance UID of the DICOM image.

The complete image curation and tagging process

Figure 4.6 illustrates the complete process used for image cohort creation, curation and annotation purposes. Using the option on the Kibana dashboard to filter DICOM image metadata stored in the Elasticsearch database, first, a high-level cohort selection solely based on the metadata is created similar to Gauriau et al. (2020). Second, an Airflow processing pipeline is triggered. The Airflow DAG, which is illustrated in Figure 4.6, takes as input the Elasticsearch query defining the cohort along with project information necessary for Doccano. The workflow creates a Doccano project with the collection of study IDS as text input using the API of the Doccano backend. Once the project is created, the user can proceed with the annotation process on the Doccano UI. A screenshot of the tagging process is provided in the results section 5.4.2 in figure 5.29. To transfer the created labels to the DICOM image, Doccano's UI was extended with a button to trigger another Airflow DAG. The workflow downloads the annotated data from Doccano and stores the created tags into the metadata of the DICOM image. In its current form, the Clinical Trial Protocol ID Attribute (0012,0020) is used to store the created tags as a comma-separated string. It is the same tag into which the Application Entity (AE) title is written, which is specified when transmitting DICOM data to the platform. Within Kaapana, the tag serves as a data identifier. However, in principle, the annotated data which is received from a text classification task or also from a free-text annotation can be stored in any suitable DICOM tag.

4.5 Real-world federated learning for the task of image segmentation

The development of robust and generalizable models requires training on a big amount of heterogeneous data. Especially in the medical domain, access to data from different institutions is often hampered by regulatory and data privacy issues, which are presented in Section 1.3.1. Federated learning is a promising strategy to circumvent a lot of those issues. This section presents a technical solution that enables federated learning scenarios between different Kaapana instances. The design of the solution is adapted so that all regulatory and data privacy requirements within the RACOON project are covered.

For federated learning applications in RACOON such as training s segmentation model, three major extensions are necessary within Kaapana. Firstly, a dedicated backend and user interface need to be implemented to manage the federated communication between the clinical sites and the central site(s). Secondly, locally running workflows need to be equipped with the possibility to share data with the central instance. Lastly, the implementation of the segmentation algorithm, here the nnU-Net, needs to be adapted to work in federated use cases (Kades et al., 2022b). All other components needed for federated learning are already present in the current Kaapana toolkit (cf. Section 3.1.1).

The implementation is evaluated by the feasibility of training the nnU-Net in a federated setup as well as the performance of the federated trained model in comparison with centralized, single-site and ensembling training approaches.

4.5.1 Backend and user interface

The implemented functionalities of the backend and the user interface are mainly based on the technical requirements and the setup of RACOON. Within RACOON a Kaapana-based platform is deployed multiple clinical sites, which are all connected to a dedicated central Kaapana-based platform. The main requirement within RACOON is that the local instances only communicate via a unidirectional network channel with the central instance, due to network security within the hospitals. As an additional security layer, on the clinical site, only a set of Uniform Resource Locators (URLs) is whitelisted for installation and update purposes as well as for communicating with a dedicated central instance. The proposed backend for federated use cases in Kaapana is responsible for transferring models between the different instances using a secure unidirectional Secure Sockets Layer (SSL) communication and for managing workflows running on the instance itself (referred to as "client instance") and on external instances (referred to as "remote instances"). The authentication between the local and the central instance is realized using custom authentication tokens. For a robust file transfer between the instances, the implementation makes use of the Kaapana integrated MinIO S3 object store. Therefore, any file transfer via the backend is forwarded to the MinIO S3 object storage of the central Kaapana instance. The object store makes use of so-called pre-signed URLs for a secure file transfer. Transferred files containing the models can be additionally protected using Fernet encryption. For the management of workflows on the client instance and remote instances, a job queuing system is implemented within the backend. To execute a job on a remote Kaapana instance, it is first queued on the central instance with the specification to be executed on a remote instance. In parallel, the remote instance periodically checks if new jobs are available at the central instance. If yes, the queued jobs are pulled and executed by the remote instance. During the job execution, regular updates of the workflow's state are reported to the central instance. Whether to periodically or manually fetch and execute new jobs can be specified on the remote instance. The control of job fetching and execution allows users of remote instances to first inspect the incoming workflow parameters before executing it. The main functionality of the implemented job queuing architecture is, therefore to execute from a client instance an Airflow workflow on a remote instance. This functionality allows, from a technical point of view, the implementation of any federated learning strategy. In the centralized federated learning setup presented here, jobs are executed periodically from the central instance on the local instances. However, the functionality is not limited to usage in federated learning scenarios, for example, it might also be used to execute only a single job on a remote instance, such as an evaluation workflow that transfers its results back to the central instance.

The Vue-based frontend (Vue.js, 2023) presents a user interface for the interaction with the backend. It provides input forms for adding the client instance and remote instances to establish a connection between them. The user at each instance can configure whether jobs should be fetched and executed automatically. Furthermore, workflows and tag-based image data which are available for an execution from a remote instance have to be specified on each client instance. Finally, the user interface provides a detailed overview of all submitted jobs with detailed information on their workflow-specific parameter and their current state.

4.5.2 Local workflow adjustments

In Kaapana, data processing pipelines (workflows) are executed using Airflow. In Airflow, a workflow is represented by a DAG, which consists of multiple building blocks called operators. An operator either executes a python script or launches a processing container on the Kubernetes cluster. By default, all workflows within Kaapana are designed to run locally. However, for federated use cases, an exchange of data is required. The here presented solution allows this data exchange for any workflow, with only little modifications to the DAG implementation itself by introducing so-called hooks to the operators. Like this, in principle, any workflow could be used for federated scenarios. In detail, to adjust local workflows for federated use cases, a configurable pre-hook to download data and a post-hook to upload data to a remote instance are added to each operator. The hooks also allow skipping the execution of an operator or loading data from a previous workflow run. This functionality enables running workflows multiple times for different purposes such as pre-processing runs, the actual training runs and possible post-processing runs. In Section 5.5.2, the usage of the introduced hooks is demonstrated on the example of the federated training of the nnU-Net. The nnU-Net workflow itself is introduced in Section 3.1.2.

The execution of federated training might be error-prone due to connection failures and the interplay between many independent systems. To allow error-free and robust training despite network issues, multiple error exceptions and retries are implemented at different levels throughout the processing pipelines and backend implementation. In addition, in case the training process stops for unexpected reasons, the training can be recovered and continued at its last successful training step.

To maintain a high level of security, the governance of the federated functionality for the workflows is controlled by the local instance. For each operator, the whether to share data can be activated, deactivated or customized to only share specific files with a central instance. Therefore, operators of an external remote instance cannot maliciously manipulate the locally running workflows.

4.5.3 The nnU-Net in federated settings

The peculiarity of the nnU-Net is that the preprocessing pipeline and the model architecture are configured individually for each training dataset using heuristic approaches based on local data characteristics. In detail, a so-called fingerprint of the training dataset is created, which is then used to configure a segmentation pipeline. For a federated training of the nnU-Net, it is therefore required to create the fingerprint based on the characteristics of all data which are distributed across the participating sites. Otherwise, federated training is not possible due to inconsistent pre-preprocessing steps and model architectures per site. In the solution proposed, the fingerprint on all data is created using a preparation round, in which the data characteristics from the local instances are shared with the central instance, which then combines and redistributes a common fingerprint to the local sites. It should be noted that the fingerprint does not contain any person-specific information to comply with data protection and regulations.

5 Experiments and results

5.1 Semantic modeling for semantic textual similarity

5.1.1 Experiment setup

All methods presented in Section 4.1 were applied to the text data from the 2019 N2C2/OHNLP Clinical STS track introduced in Section 3.2.1 consisting of 1642 training and 412 test sentence pairs. This section presents the experimental setup for the actual training and testing of the presented methods, including preprocessing steps, implementation details for the different approaches, a dataset analysis, and details on the evaluation runs.

Preprocessing

Before applying the methods, a set of preprocessing steps is applied to the data to unify and clean the dataset. Depending on the use case a subset of the following preprocessing steps are suitable:

- ContractionExpander: To normalize the text and reduce ambiguities a contraction expansion (e.g. "we'll" → "we will") using the pycontraction Python package (PyContractions, 2023) is done.

- NumberUnifier: To unify the representation of words that convey the same meaning, all textual representations of numbers are converted to the corresponding numerical literal (e.g. "forty-two" → "42").

- SpellingCorrector: Spelling mistakes in the datasets are corrected (e.g. "refil" → "refill")
- MedicationRemover: In preparation for the medication graph all vendor drug names are removed from the medication sentences and only the general active agent names are kept (e.g. "metoprolol succinate [TOPROL XL] 25 mg..." → "metoprolol succinate 25 mg...")
- SentenceTokenizer: Before the application of a word tokenizer, sentences are split using the library segtok (segtok, 2023). The library splits a sentence on common sentence markers (e.g. ".,?") and then evaluates again every split by considering the surroundings and checking for false positives (e.g. to handle cases like name initials correctly.)
- WordTokenizer: After the sentence tokenizer, words are tokenized using the word tokenizer from the library segtok.
- PunctationRemover: Punctations are removed using a rule-based approach
- LowerCaseTransformer: Uppercase letters are transferred to lowercase letters
- StopWordsRemover: Stop words, as well as task-specific words with a high frequency such as "tablet" or "medication", are removed to increase the variety of sentences for the similarity measures.
- Lemmatizer: Words are normalized (e.g. "moved steadily" → "move steadily") using the python library pattern (De Smedt and Daelemans, 2012; pattern, 2023).

From the introduced preprocessing steps, the ContractionExpander, NumberUnifier, SpellingCorrector, and LowerCaseTransformer are applied before the use of BERT, and all steps except the MedicationRemover are applied when calculating the features set 1 and 2.

Feature sets and voting regressor for the EnhancedBERT

The two Feature sets used for the EnhancedBERT are created by empirically evaluating combinations of similarity measures during the development.

Feature set 1, which is added to the CLS token, consists of the following token-based text distance measures with n-grams of $n = 3$ (i.e., three characters are used for comparison): Jaro and Jaro-Winkler distance. Sørensen-Dice coefficient, overlap coefficient and cosine similarity. In addition, the InferSent2 sentence embeddings are calculated as well as a mean pooled sentence representation of the GloVe word embeddings (used model: glove.840B.300d (GloVe, 2023)). Different distance metrics are calculated for each embedding of the corresponding sentence pairs: Cosine similarity, Euclidean, Manhatten and Minkowski distance.

Feature set 2 consists of text distance measures with n-grams of $n = 3$ and $n = 4$: Damerau-Levenshtein, Jaro Winkler, the Bag distance, and a square root-based normal compressor distance (SqrtNCD). Furthermore, the Cosine similarity, Euclidean, Manhatten and Minkowski distance are calculated, between the Inferset1 embeddings of the sentence pairs and between the GloVe word embeddings of the sentence pairs.

The used Voting Regression consists of a combination of linear regression models (least squares, lasso, epsilon-insensitive fitting, SVR) and ensembling models (random forest, AdaBoost, gradient tree boosting).

Medication graph training

In the medication graph approach, the learning of λ_k and the optimization of the SVR are alternated during training. To train λ_k, experiments showed that 50 update steps are sufficient to reach a convergence of the weights. The hyperparameters of the SVR model are optimized using a grid search.

Dataset analysis

To learn about possible imbalances or peculiarities of the training and test data, to understand possible shortcomings of the BERT baseline and for a more thorough interpretation of the results from the different approaches some properties of the training and test dataset are analyzed and visualized.

In the first step, the average and standard deviation (STD) of similarity scores of the training and test set are considered and the label distribution is plotted. In the second step, the number of words per dataset is plotted. Additionally, the InferSent embeddings of the training and test dataset are calculated. Using a t-SNE plot, the embeddings of training and test set can be visually compared.

Furthermore, as pointed out in the cluster analysis Section 4.1.1, weaknesses of the BERT baseline are determined by applying a k-means clustering algorithm to the InferSent embeddings with a subsequent analysis of the clusters. Within each cluster, the mean and STD of the absolute differences between the reference annotations and the BERT predicted similarity scores of all sentence pairs are calculated. The differences are visualized in the next section in a t-SNE plot and a box plot.

Evaluation runs

The methods are evaluated on the dataset from the 2019 N2C2/OHNLP Clinical STS track. The test set consists of 412 sentence pairs and the training set of 1642 sentence pairs. To increase the comparability of the models and to reduce noise in the data a

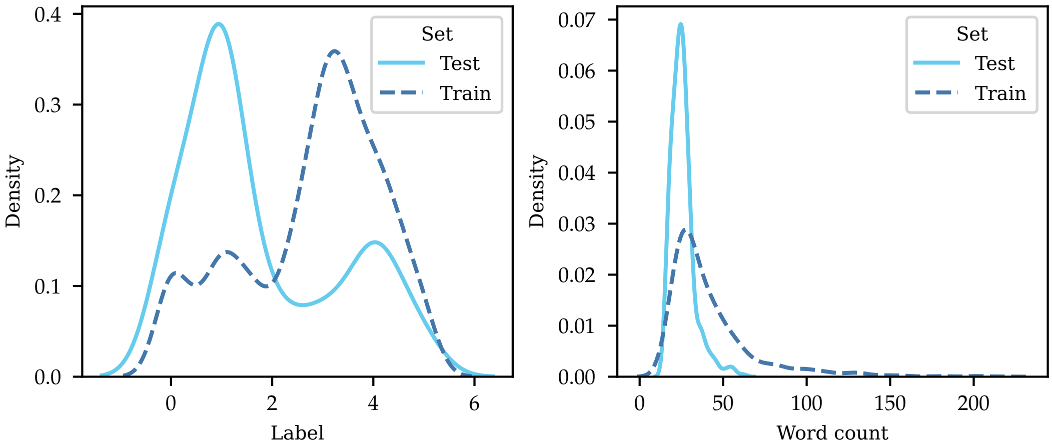


Figure 5.1: *Distribution of labels and number of words for the training and test dataset. This figure was adapted from Kades et al. (2021).*

k =150 cross-fold validation is applied. By concatenating the results on the nth-folds, a validation set is created. In the following, both terms, training and validation set, are used to describe the same set of 1642 sentence pairs. The evaluation metric for all methods is the PCC.

In all approaches, the pre-trained ClinicalBERT (Alsentzer et al., 2019)[1], which builds upon BioBERT (Lee et al., 2020), is fine-tuned for the task of semantic textual similarity. Using the PyTorch HuggingFace Transformers library (Wolf et al., 2020), the training is executed for 10 epochs with a maximal sequence length of 128, a learning rate of $2^{-}5$. After training, the model with the lowest loss on the validation set is used for testing.

Using k-fold cross-fold validation, the PCCs for the validation set are calculated on the predicted similarity scores of the concatenated nth-folds. For the test set, an additional ensembling technique is employed. In detail, with the trained models of the different folds, separate predictions are created and then averaged for each sentence pair in the test set before calculating the PCC. A consequence of the employed approaches to create the final similarity scores for each sentence pair is that no information about the variance can be given because only one PCC per set is calculated.

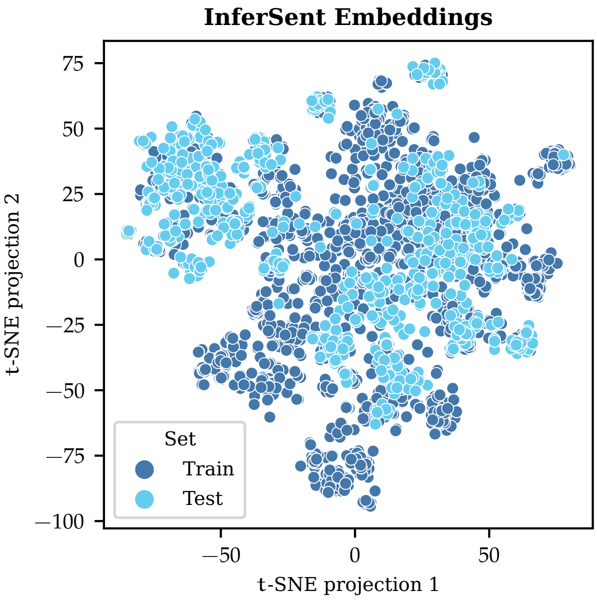


Figure 5.2: *Visualizations of the training and test dataset using t-SNE projected InferSent embeddings. Each point corresponds to a sentence pair of the training or test dataset. The graph showcases that the types of sentences in the test set are only a subset of the types of sentences in the training set. This figure was adapted from Kades et al. (2021). t-SNE=t-distributed stochastic neighbor embedding.*

5.1.2 Results

Dataset evaluation

The analysis of the training and test dataset reveal some shift between the training and test dataset. Therefore, the average similarity score in the training set of around 2.79 is higher than the average similarity score in the test dataset of around 1.76. However, the STD in the training set of around 1.39 is slightly smaller than in the test set with 1.52. For visual perception, the distribution of labels is illustrated on the left in Figure 5.1. It shows that similarity scores of around one are represented most in the test set, whereas similarity scores of around three are the most prominent in the training set. Also, the distribution of word counts per dataset exhibits discrepancies between the training and test dataset. Therefore, with an average length of around 26±7 words per sentence pair, the sentence pairs in the test set are shorter than the sentences in the training set with an average length of around 42±26 words per sentence pair. On the right in Figure 5.1 also the distribution of word counts per

---

[1] Name of the pre-trained ClinicalBERT model: biobert_pretrain_output_all_notes_150000

Approach	Validation set	Test set
Approach 0: Baseline		
ClinicalBERT	0.850	0.859
Approach 1: Voting Regression		
Enhanced BERT	0.851	0.859
Voting Regression	0.860	0.849
Approach 2: M-Heads		
Enhanced BERT with M-Heads	0.853	0.876
Enhanced BERT with M-Heads + Med. graph	0.853	**0.883**
Approach 3: Medication graph		
Voting Regression + Med. graph	**0.862**	0.862

Table 5.1: *The PCC for the validation and test set of the different approaches, which are presented in detail in Figure 4.1. The results are rounded to 3 decimal places and the best results are printed in bold. This table was adapted from Kades et al. (2021). PCC=Pearson correlation coefficient.*

dataset is plotted. In addition, the InferSent embeddings illustrated in Figure 5.2 color-coded with the corresponding set suggest that the sentence pairs in the test set only represent a subset of those occurring in the training set. For example, the slightly separated clusters in blue have no test set embeddings in their direct neighborhood. In Figure 5.3 the InferSent embeddings of the sentence pairs color coded with the corresponding cluster from the k-means clustering are illustrated. The opacity of the points corresponds to the absolute difference between the BERT predicted similarity scores and the reference annotations (the stronger the color the higher the deviation). Additionally, the mean and SD of each cluster are reported in the legend. Figure 5.4 illustrates the distribution of the absolute differences per cluster along with the number of samples per cluster, suggesting that for sentence pairs belonging to cluster 3, BERT struggles the most. As mentioned in the methods Section 4.1.1, cluster 3 is dominated by sentences that prescribe medication and therefore motivates the medication graph approach.

Evaluation runs

In Table 5.1, the PCC scores of the different developed approaches and combinations on the validation and test sets are presented. A combination of EnhancedBERT, M-heads and the medication graph yields the best PCC of 0.883. In comparison, a PCC of 0.901 was achieved by the winner of the 2019 N2C2/OHNLP Clinical STS track from IBM Research reached.

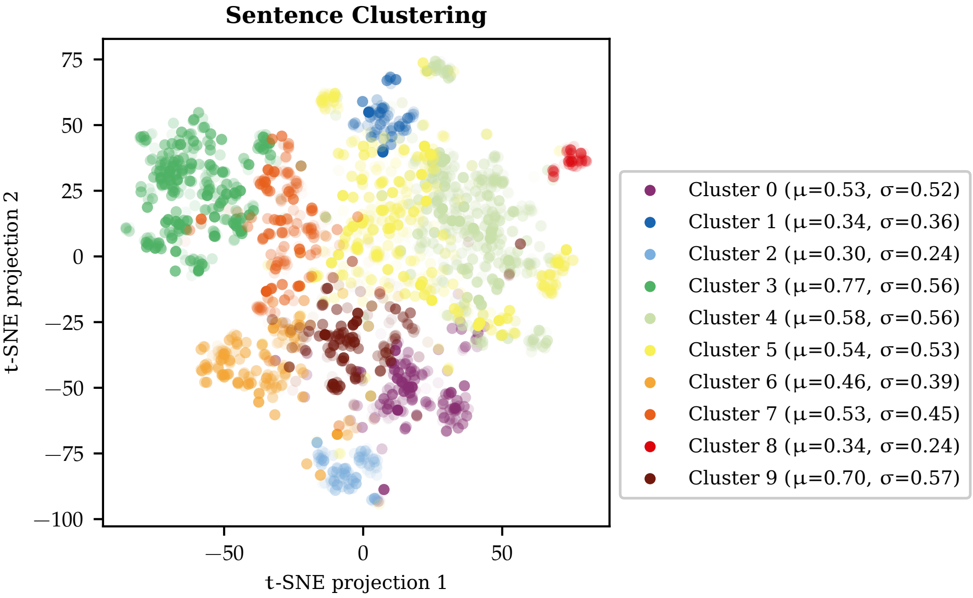


Figure 5.3: *Visualizations of the clusters, emerging from the k-means clustering, using t-SNE projected InferSent embeddings. Each point corresponds to a sentence pair of the training set. The analysis aims to pinpoint types of sentence pairs for which BERT points out weaknesses. The opacity of each point corresponds to the deviation between the reference similarity and the BERT predicted similarity. Therefore, the more opaque a point is, the higher the deviation from the reference. For each cluster, the average absolute difference along with the STD to the reference similarity is reported in the legend. Cluster 3, which contains sentences prescribing medications, shows the highest average differences to the reference similarities. This figure was adapted from Kades et al. (2021). BERT=Bidirectional Encoder Representation from Transformers, STD=standard deviation, t-SNE=t-distributed stochastic neighbor embedding.*

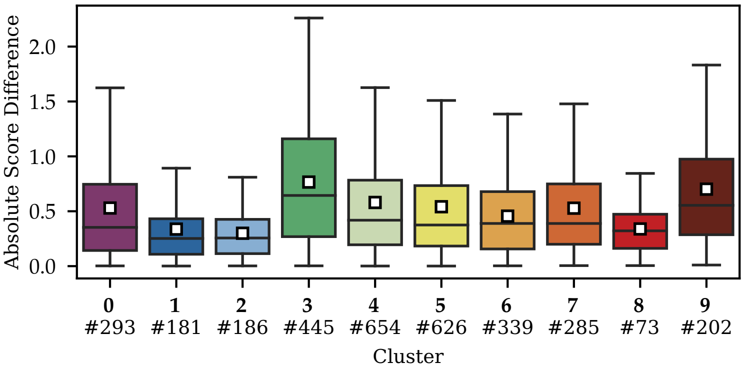


Figure 5.4: *Illustration of the difference between the reference annotations and the BERT predicted scores along with the number of sentence pairs per cluster below the cluster number. For each cluster the lower and upper quartile, and the median position is illustrated, emphasizing the opacity information from Figure 5.3. The whiskers show the high variations within each cluster and the white square denotes the mean value. This figure was adapted from Kades et al. (2021). BERT=Bidirectional Encoder Representation from Transformers.*

5.2 Semi-structured data analysis for text summarization

5.2.1 Experiment setup

All approaches presented in Section 4.2 are applied on the SORs from the UKHD introduced in Section 3.2.2. Particular challenges arise when working with the SORs for the task of text summarization. Those include the segmentation of sentences due to the custom structure of the report and the clinical jargon as well as imbalances in the data. The imbalance has its origin in the TRC of the reports, e.g. findings indicating a PD are more likely to be complex than reports indicating a CR. For this reason, the TRC is also taken into account in the evaluation of the approaches. This section presents details on the experimental setup including preprocessing steps and implementation details on the evaluation runs.

Preprocessing steps

As a preprocessing step, the sentences in each section are segmented. The segmented sentences are used when extracting key sentences from the findings section (cf. Section 4.2.1) and for the construction of structured sequences (cf. Section 4.2.1). By design, the SORs are structured into sections and blocks, in which each block contains free-form narrative text snippets. Due to the fixed structure that is described in Section 3.2.2 in more detail, a tailored sentence segmentation approach for the sections is applied. In the first step, the main blocks, the general section and the finding sections including the primary tumor location, metastases, reference measurements and non-oncology findings are segmented. In the second step, the sentences of the blocks are segmented using a custom regex, which is tailored to take into account report-specific cases such as abbreviations, dates, and serial numbers, in which a separation based on a period would be wrong. Whereas this is more complex for the findings section, the general section normally consists of only two sentences describing the treatment situation and the previous examinations.

Evaluation runs

In total, 10514 SORs from the years 2018 and 2019 are split into a training (80%), validation (10%) and test (10%) set to evaluate the different text summarization approaches. To evaluate the significance of the sentence extraction method for the models including the extractive task, runs are included in which the extracted key sentences are chosen randomly from the finding sections.

Again, all presented implementations make use of the PyTorch HuggingFace Transformers library by Wolf et al. (2020). For training the models, 8410 reports serve as training reports and 1052 as validation samples. The remaining 1052 reports are reserved for testing. The model is trained for 10 epochs with an early stopping such that the training is terminated if the validation loss is no longer decreasing within 3 epochs.

Before the actual training of the models, experiments are executed to determine the most effective methods to extract key sentences from the source which are presented in Section 4.2.1. The extraction approaches are evaluated based on the recall sore of the ROUGE−1 and ROUGE−L metrics because the extracted sentence should overlap as much as possible with the reference conclusion, independent of the number of false positives.

Automatic evaluation

The evaluation metric for the text summarization methods is the F1 score of the ROUGE−1 and ROUGE−L metric (Lin, 2004). ROUGE measures the token overlaps between the reference and generated conclusion. Besides an overall score, the scores are also reported in dependency of the different TRCs. In addition as a way to measure the factual correctness, it is examined, whether the predicted TRC corresponds to the one in the reference conclusion.

Human evaluation

Also, a human evaluation is applied to a subset of the test data based on comprehensibility as well as oncological and non-oncological correctness. Similar to the automated evaluation the human evaluation is done in dependency of different TRCs. More details on the human evaluation are given in Section 4.2.2. A screenshot of the annotation process in Doccano is given in the appendix in Figure B.2.

To get an impression of how significant the differences in the generated conclusions are during human evaluation, a rank-based evaluation is also applied. In detail, using the Likert-scaled scores from the human annotators a ranking list for each summarized report and each evaluation criteria is created. The results are visualized in the next section.

5.2.2 Results

Table 5.3 reports the recall scores of the ROUGE score for the different sentence extraction approaches. The Longest-k method yields the highest agreement between the extracted sentence and the conclusion.

In Table 5.2 the average number of sentences per section resulting from the sentence segmentation, the number of samples and the TRC count per dataset is reported.

The ROUGE-1 F1 scores for the presented text summarization methods as well as the random variations are presented in Table 5.4. All hybrid models outperform the BERT2BERT baseline by a clear margin. Overall, the BERT2BERT+Ext+Ptr creates

Section / Set	Training (# 8410)	Valid (# 1052)	Test (# 1052)
General	2.0	2.0	2.0
Findings	21.1 ± 8.2	20.5 ± 7.5	21.7 ± 7.5
Conclusion	3.1 ± 2.0	3.4 ± 2.0	3.5 ± 2.0
TRC			
CR	2970	168	207
PR	668	92	77
SD	2998	496	476
PD	1774	296	292

Table 5.2: *Number of samples for the training, validation and test set as well as the average number of sentences per section after sentence segmentation and the TRC counts per dataset. The general information section consists always of two sentences containing background information, the finding sections of around 22 sentences with a variation of 8 sentences and the conclusion of around 3 sentences with a variation of 2 sentences. This table was adapted from Liang et al. (2022). CR=complete response, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

the best conclusions, however, with only small improvements in comparison to the BERT2BERT+Ext and BERT2BERT+Ptr. The reports with the CR class are handled well by all models. The variations with randomly selected target sentences for the extraction approach perform inferior to the ones with Longest-k, however, still superior to the BERT2BERT baseline.

The accuracies of the factual correctness check are reported in Table 5.5. Except for the PR class, all accuracies are above 0.7.

Human evaluation

The average scores awarded by the two annotators in the human evaluation for the three criteria comprehensibility and oncological and non-oncological correctness are plotted per TRC in Figure 5.5. The ranking results are illustrated in Figure 5.7 and Figure 5.6. The first one shows the rank stability over different criteria and TRC and the second one shows the rank distribution for the different criteria and the overall ranking. The ranking results demonstrate that in the human evaluation, all hybrid methods perform relatively equally and in the case of the comprehensibility criteria even better than the reference. The stability of the rank indicates that the reference represents an upper baseline and the BERT2BERT a lower baseline in the human evaluation. However, there are also fluctuations visible in the criteria and the TRCs.

In Table 5.6 a full SOR is illustrated along with the reference conclusion. The corre-

Meric	Longest-k	Tfidf-Ex	TextRank
Rouge-1	**41.9**	40.6	40.4
Rouge-L	**40.8**	38.7	39.6

Table 5.3: *Recall scores of the ROUGE metrics for the three key sentence extraction approaches. With the score, the overlap between the references and the extracted key sentences is measured. The extraction always included the two sentences from the general information section. The Longest-*k *method is evaluated using* $k = 4$ *since the reference conclusion does not exceed six sentences. This table was adapted from Liang et al. (2022). ROUGE=Recall-Oriented Understudy for Gisting Evaluation, TF-IDF=term frequency-inverse document frequency.*

Method	whole	CR	PR	SD	PD
BERT2BERT	36.15	55.27	30.86	32.09	30.93
BERT2BERT+Ext	42.13	**58.99**	38.19	38.17	36.68
BERT2BERT+Ext (random)	37.27	57.22	31.43	32.71	31.32
BERT2BERT+Ptr	42.25	55.9	38.66	**39.88**	**39.04**
BERT2BERT+Ext+Ptr	**43.32**	57.91	**40.15**	39.39	38.65
BERT2BERT+Ext+Ptr (random)	42.1	57.37	38.71	39.41	37.81

Table 5.4: *ROUGE-*1 *F1 scores for the different summarization methods on the test as well as on the subsets of the four TRC. For the BERT2BERT+Ext(+Ptr) methods, scores are reported in which the sentences are extracted randomly instead of using the Longest-*k *method. This table was adapted from Liang et al. (2022). BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, PD=progressive disease, PR=partial response, ROUGE=Recall-Oriented Understudy for Gisting Evaluation, SD=stable disease, TRC=tumor response category.*

sponding generated conclusions of the models with, again, the reference conclusion are given in Table 5.7.

Method	CR	PR	SD	PD
BERT2BERT	0.78	0.40	0.73	0.75
BERT2BERT+Ext	0.77	0.52	0.72	0.75
BERT2BERT+Ptr	0.78	0.22	0.71	0.75
BERT2BERT+Ext+Ptr	0.76	0.62	0.73	0.79

Table 5.5: *Accuracies for the summarization methods and different TRC. This table was adapted from Liang (2021). CR=complete response, PD=progressive disease, PR=partial response, SD=stable disease.*

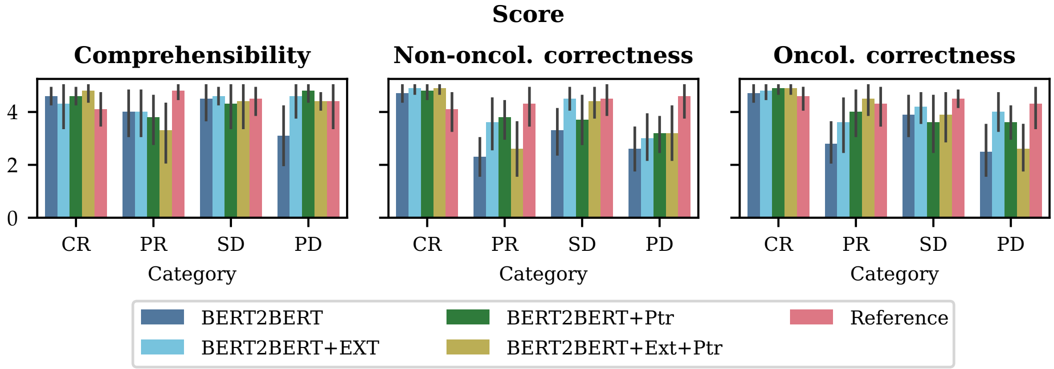


Figure 5.5: *Scores assigned during the human evaluation for the reference summaries and the summaries generated by the different presented methods. The average scores with their STDs are plotted for the three evaluation criteria and respectively for each TRC. This figure was adapted from Liang et al. (2022). BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, PD=progressive disease, PR=partial response, SD=stable disease.*

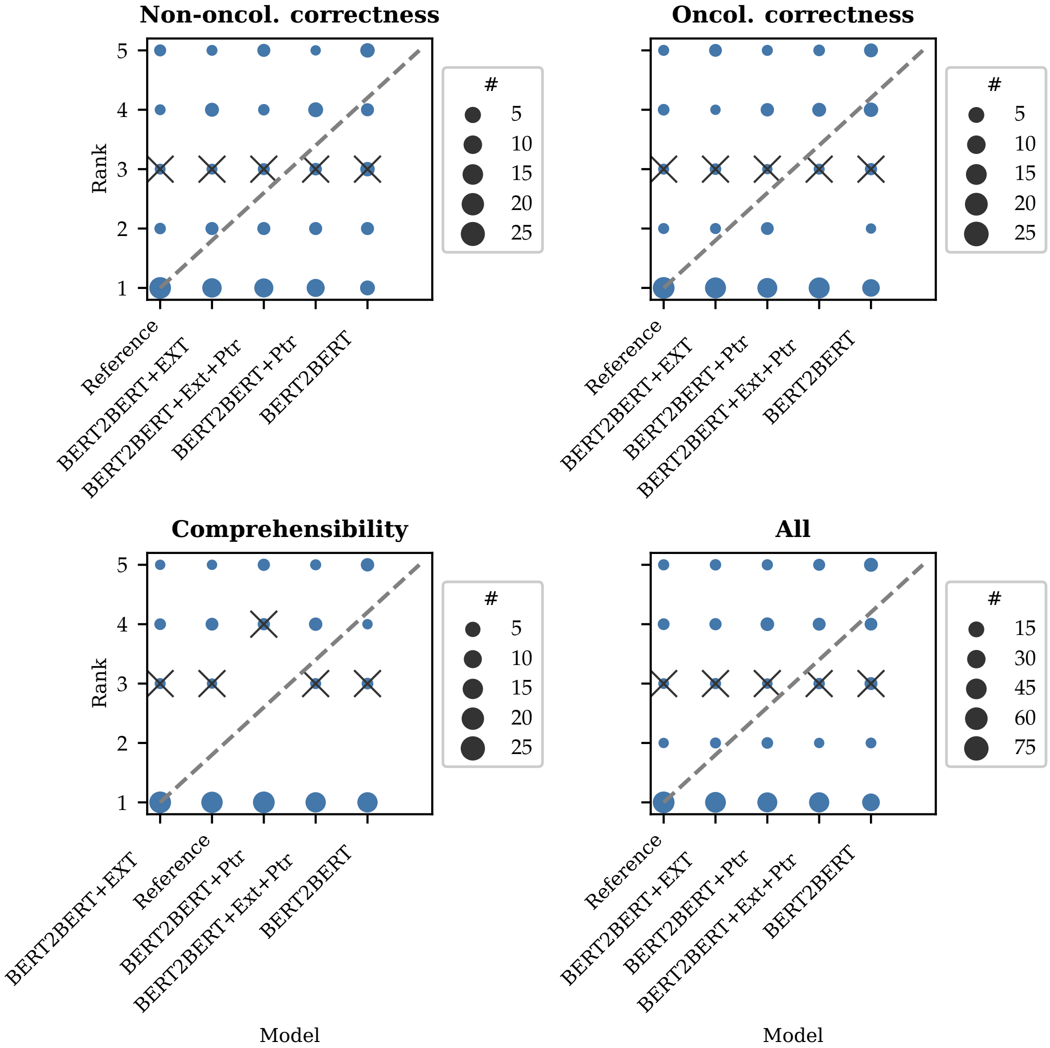


Figure 5.6: *Rank distribution resulting from the human evaluation for all samples and, respectively, for each of the three criteria: non-oncological and oncological correctness and comprehensibility. Visualized is the mean rank of the reference conclusion and the four generated conclusions. The rank is calculated based on assigned scores. The best conclusion receives the lowest rank. In case multiple conclusions receive the same score, they all get the same best rank. The size of the points indicates the number of times a conclusion reached a rank. The "X" indicates the median rank. The distribution shows that there is little difference in performance in the human evaluation, with the reference still outperforming the generated conclusions overall. BERT=Bidirectional Encoder Representation from Transformers.*

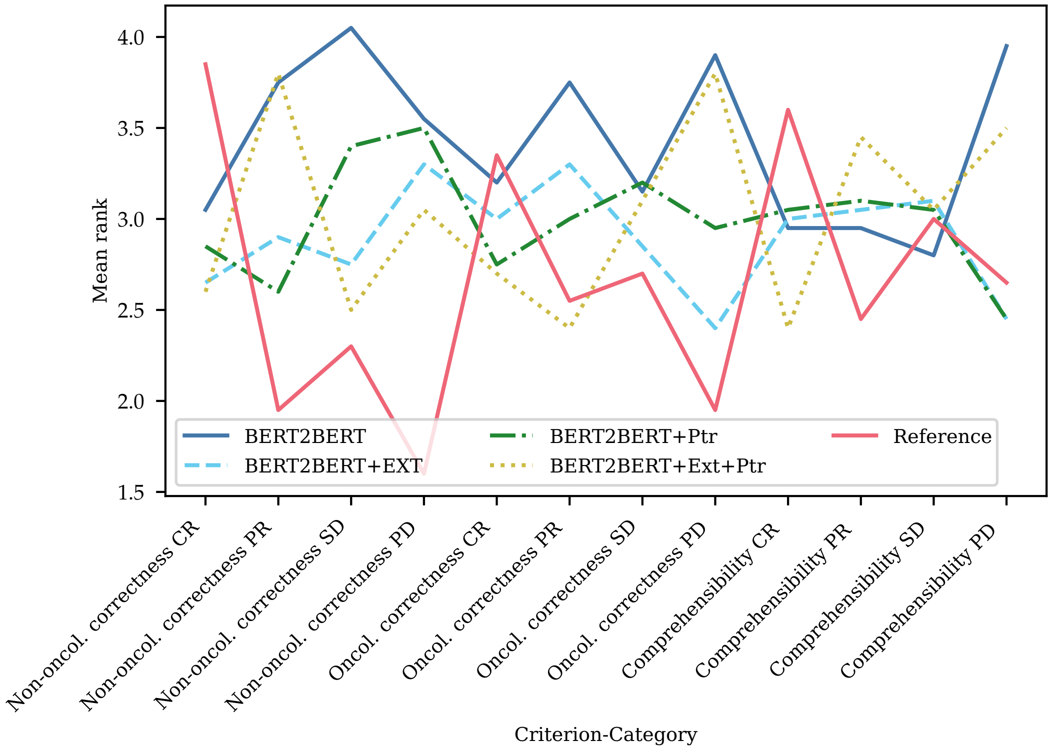


Figure 5.7: *Visualization of the mean rank resulting from the human evaluation across the three criteria and the four TRCs. The graph illustrates variations between the generated conclusions, however, with tendencies of the references representing the upper baseline and the BERT2BERT model the lower baseline. BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

Section	Content
General	Untersuchungsregion Thorax (CT), Abdomen (CT) Behandlungssituation Ausgangsbefund. Vergleich Letzte Vergleichsuntersuchung: 05.07.2018.
Findings	Primärtumor / Lokalrezidiv Soweit messtechnisch erschwert erfassbar progrediente diffus infiltrierende Raumforderung des Pankreaskopfs mit Gangstau im Pankreasschwanz und vollständiger Ummauerung des Truncus coeliacus, mindenstens 180ř Ummauerung der A. liniealis . Bekannter kompletter Verschluss der extrahepatischen Pfortader und V. mesenteria superior mit ausgeprägten Kollateralen . Regionäre Lymphknoten Gering prominenterer vermehrter Lymphknotenbesatz mesenterial, exemplarisch mit einem KAD von 7 mm, zuvor 5 mm (8-137) . Metastasen Lunge und Pleura: Keine . Thorakale Lymphknoten und Weichteile: Keine . Leber: Keine . Abdominale Lymphknoten und Weichteile: Keine . Peritoneum: Kein eindeutiger Nachweis einer Peritonealkarzinose, jedoch Infiltration der Mesenterialwurzel durch den Primarius und Nachweis geringer freier Flüssigkeit im kleinen Becken . Skelett: Keine . Referenzmessungen Keine Referenzmessungen . Thorax Portsystem in situ . Kein katheterassoziierter Thrombus . Abdomen Unveränderte Nebenmilze . Konstante geringe Cholestase Skelett Im Übrigen ohne besondere Auffälligkeiten .
Conclusion	Onkologische Beurteilung Onkologisch Befundkonstanz mit tendenzieller Zunahme der Tumorlast gegenüber 05.07.2018. Progredient diffus infiltrierende Raumforderung des Pankreaskopfs. Andere Befunde Keine Angabe.(Progressive diffusely infiltrating mass of the pancreatic head.)

Table 5.6: *Anonymized sample SOR with the general information section and the findings section. The reference conclusion and generated conclusions are given in the next table. This table was adapted from Liang et al. (2022).*

Methods	Generated conclusion
Reference	Onkologische Beurteilung Onkologisch Befundkonstanz mit tendenzieller Zunahme der Tumorlast gegenüber 05.07.2018. Progredient diffus infiltrierende Raumforderung des Pankreaskopfs. Andere Befunde Keine Angabe.(Progressive diffusely infiltrating mass of the pancreatic head.)
BERT2BERT	Onkologische Beurteilung Onkologisch Befundverschlechterung gegenüber dem Ausgangsbefund vom 11.04.2018 und 12.05.2019. In diesem Zeitraum zunehmende Peritonealkarzinose mit Ummauerung der VMS, sowie Ummauerung des TCs bei bekanntem Magenkarzinom. Andere Befunde Keine Angabe.(During this period, increasing peritoneal carcinomatosis with walling of the VMS, as well as walling of the TC in known gastric carcinoma.)
BERT2BERT+ Ext	Onkologische Beurteilung Onkologisch Befundkonstanz gegenüber 05.07.2018. In diesem Zeitraum keine wesentliche Änderung des Pankreasschwanzes mit Infiltration der Pfortader und Ummauerung der Truncus coeliacus, AMS / VMS sowie Arteria hepatica communis durch die AMS - suspekte Lymphknotenmetastasen. Andere Befunde Keine Angabe.(During this period, no significant change in the tail of the pancreas with infiltration of the portal vein and walled truncus coeliacus, AMS / VMS and arteria hepatica communis by the AMS - suspicious lymph node metastases.)
BERT2BERT+ Ptr	Onkologische Beurteilung Onkologische Befundverschlechterung zum 05.07.2018. In diesem Zeitraum neu aufgetretene Lymphknotenmetastase im Oberbauch mit ausgeprägter Infiltration der Pfortader und der V. Pankreatitis. Unveränderte Darstellung der bekannten Peritonealkarzinose. Andere Befunde Keine Angabe.(Newly appeared lymph node metastasis in the upper abdomen during this period with marked infiltration of the portal vein and pancreatic vein. Unchanged presentation of known peritoneal carcinomatosis.)
BERT2BERT+ Ext+Ptr	Onkologische Beurteilung Onkologisch Befundkonstanz mit tendenzieller Zunahme der Tumorlast gegenüber 05.07.2018. In diesem Zeitraum moderate Zunahme des Primärtumors im Pankreasschwanz. Bekannter Verschluss der extrahepatischen Pfortader und Ummauerung der A. gastrica sinistra und Infiltration der Pfortader. Andere Befunde Keine Angabe.(Moderate increase in primary tumor in the pancreatic tail during this period. Known occlusion of the extrahepatic portal vein and walling of the gastrica sinistra artery and infiltration of the portal vein.)

Table 5.7: *Reference conclusions and conclusions generated by the four text summarization methods. The red marked parts for the BERT2BERT show randomly generated dates. In contrast, the extraction approach and the pointer mechanism generated the correct date. The green marked parts indicate phrases extracted from the finding section of the report. This table was adapted from Liang et al. (2022). BERT=Bidirectional Encoder Representation from Transformers.*

5.3 Assessing distributional shifts for text classification

5.3.1 Experiment setup

The performances of TRC classification methods presented in Section 4.3 are applied to the German radiology reports presented in Section 3.2.2. This section presents details on the data acquisition as well as the generation of the medical and technical characteristics of the different corpora. Furthermore, besides information on the reference annotation and human baseline generation, implementation details on the evaluation runs, the hyperparameter optimization and the visualization of results are given.

The report corpora

As described in Section 3.2.2, the original report corpora retrieved from the UKHD, DKFZ and TKH consist of 14569 reports. Figure 5.8 illustrates in a flowchart the creation of training and test sets. From the original 13685 SORs, 852 duplicate reports are removed and another 3180 are omitted because of difficulties to extract the TRC category using the regex. The remaining 9653 SORs are split into a training set consisting of 8653 and a test set consisting of 1000 SORs. From the 412 FTORs of the DKFZ and 472 of the TKH 369 and 433 reports, respectively, created the test set because some reports are removed due to missing evidence of cancer in patients' radiologic history or no clear assessment of tumor burden change using short- and long-term imaging (Fink et al., 2022a).

For the annotations and the application of algorithms, the reports are curated and preprocessed. Similar to the preprocessing steps in section 5.2.1 the reports are first split into the three main blocks "general information", "findings" and "conclusion" using a regex. Furthermore, also with the help of a regex, the TRC from the conclusion of the SORs is extracted. Patient characteristics like the exam date, age, the number of visits per patient and sex are given as metadata. In case the sex is not given it is determined using the library gender-guesser (Gender Guesser, 2023). Additionally, using a regular expression and a hand-crafted list of diagnosis, the tumor kind and tumor family is determined. Further preprocessing steps for the algorithms include the removal of new lines and tabs as well as the tokenization of words, which is necessary for the TF-IDF-based approaches. For the BERT baseline, the tokenization is done using the BERT tokenizer, which splits the report into the smallest unit of strings that are known to its vocabulary. For the TF-IDF-based approaches, the reports are also tokenized using the BERT tokenizer for consistency reasons, however, coherent snippets are joined back together to the actual word. In the resulting sequence, punctuation is removed and words are stemmed. The removal of stop words is a hyperparameter for the TF-IDF-based approaches.

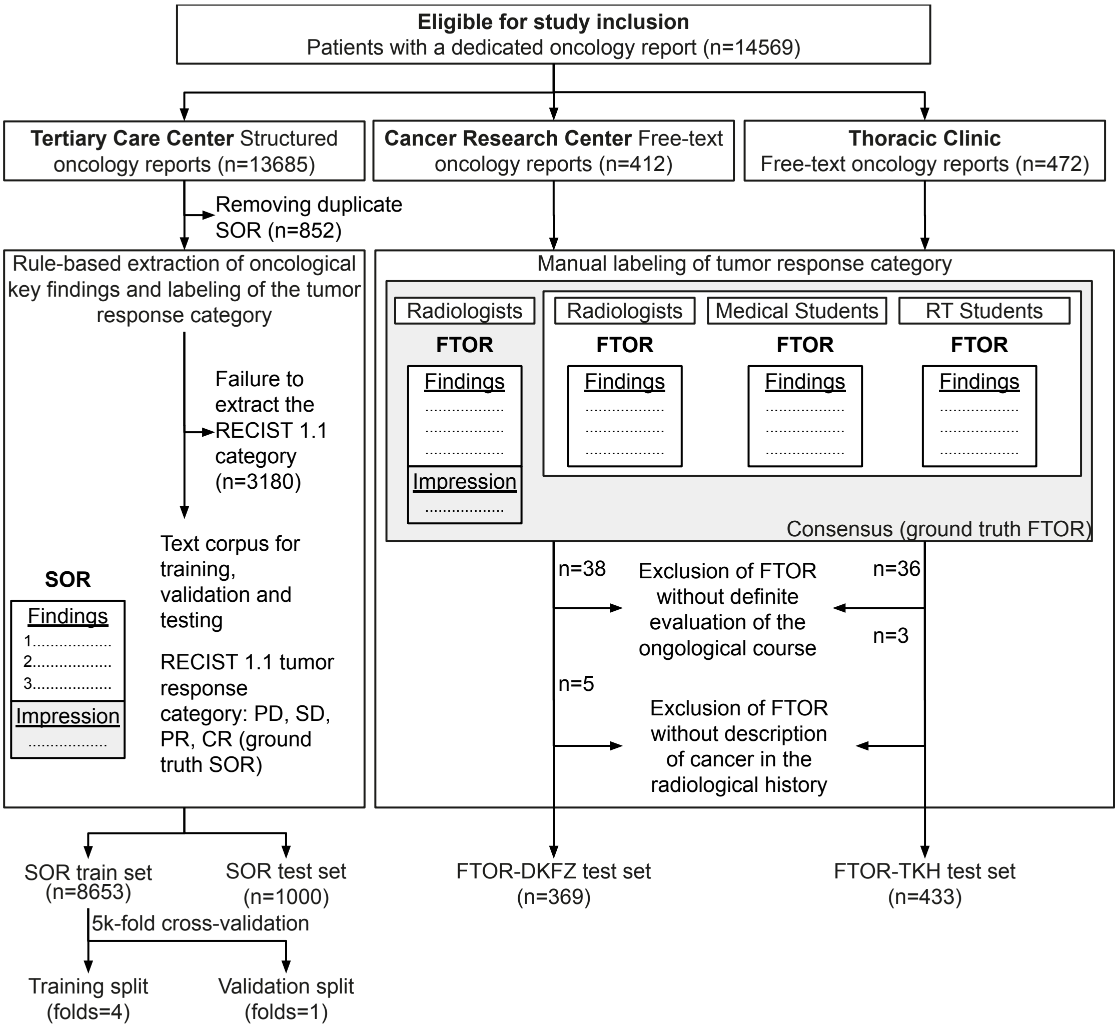

Eligible for study inclusion
Patients with a dedicated oncology report (n=14569)
Tertiary Care Center Structured oncology reports (n=13685)
Cancer Research Center Free-text oncology reports (n=412)
Thoracic Clinic Free-text oncology reports (n=472)
Removing duplicate SOR (n=852)
Rule-based extraction of oncological key findings and labeling of the tumor response category
Failure to extract the RECIST 1.1 category (n=3180)
Text corpus for training, validation and testing
RECIST 1.1 tumor response category: PD, SD, PR, CR (ground truth SOR)
SOR
Findings
1.
2.
3.
Impression
Manual labeling of tumor response category
Radiologists
Radiologists
Medical Students
RT Students
FTOR
Findings
Impression
Consensus (ground truth FTOR)
n=38
n=36
n=3
n=5
Exclusion of FTOR without definite evaluation of the ongological course
Exclusion of FTOR without description of cancer in the radiological history
SOR train set (n=8653)
SOR test set (n=1000)
5k-fold cross-validation
Training split (folds=4)
Validation split (folds=1)
FTOR-DKFZ test set (n=369)
FTOR-TKH test set (n=433)

Figure 5.8: *Flowchart illustrating the study design and the creation of the training and test sets. This figure was adapted from Fink et al. (2022a). FTOR=free-text-oncology report, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, RECIST=Response evaluation criteria in solid tumors, SOR=Structured Oncology Report.*

As presented in Section 4.3.1, next to the medical characteristics of the data, the lexical complexity of the reports is calculated. The complexity is computed for each set. To get a feeling for the complexity of the medical reports, the same analysis is applied to three publicly available datasets: WikiLingua (Faisal Ladhak and McKeown, 2020) (#58341), 10k German news articles (Schabus et al., 2017) (#10273) and Swiss Judgement Predictions (Niklaus et al., 2021) (#45183).

Human annotations

As presented in Sections 4.3.2, reference labels of TRCs for the DKFZ and TKH dataset were created by two radiologists using the annotation toolkit. For the annotation, the radiologists were presented with the complete FTOR.

Furthermore, as presented in Section 4.3.3 a human baseline for the classification of TRC along with a confidence score from a five-point Likert scale was created by seven annotators with different levels of expertise. This time only the findings section was presented to the annotators. Screenshots for the creation of the baseline annotations are presented in the Appendix B.1. In the following, the seven annotators with different levels of expertise are considered within three groups: Radiologists (Radiolgist 1 and Radiolgist 2), Medical students (Student 1 and Student 1) and RTs (RT 1, RT 2, RT 3). The agreement among the human readers for the TRC classification is reported using the inter-rater reliability metric Cohen's kappa κ, which is defined as follows (Artstein and Poesio, 2008):

$$\kappa = \frac{A_o - A_e}{1 - A_e} \tag{5.1}$$

with A_o the observed agreement and A_e the expected agreement. Cf. Artstein and Poesio (2008) for more information. Besides Cohen's kappa scores, also confusion matrices for the reference labels are presented in the results, Section 5.3.2.

NLP models

The presented BERT-based method as well as the three feature-based methods are trained on the 8653 SORs from the UKHD and then evaluated on three test datasets from UKHD, 1000 test samples, DKFZ 369 test samples and TKH 433 test samples. During training and testing, the models take as input only the findings section of the reports. The BERT model is implemented using the PyTorch HuggingFace Transformers library (Wolf et al., 2020) and the feature-based methods are implemented using the scikit-learn library (Pedregosa et al., 2018). As elaborated in Section 4.3.3, the BERT model, initialized with the pre-trained weights of the bert-base-german-cased pre-trained weights (deepset, 2023), are fine-tuned using a weighted cross-entropy on the logits resulting from the linear classification layer. To increase the generalizability and robustness of the algorithms, all models are trained using a 5-fold ($k=5$) cross-validation. In detail, for the validation performance, the results of the nth folds are concatenated to form the validation set consisting of a list of probabilities for each TRC and sample. During testing each of the five folds creates a prediction and the resulting probabilities are ensembled (averaged). As described in Section 4.3.4 it is ensured that the predicted probabilities are calibrated before the predicted TRCs are determined using the class with the maximum probability. For the generated probabilities from the BERT model, a temperature scaling introduced by Guo et al. (2017) is applied and for the feature-based methods, the probabilities are calibrated using the build-in "CalibratedClassifierCV" (scikit-learn, 2023) from scikit-learn (Pedregosa et al., 2011). To find the best model settings, an extensive hyperparameter optimization using the framework Optuna is performed (Akiba et al., 2019; Optuna, 2023). In the hyperparameter search, in total, 25 runs are executed for the BERT baseline and 250 runs for the three feature-based methods due to the much shorter training time. The tuned hyperparameter along with the best settings are illustrated in Table 5.8. A visual inspection of the hyperparameter optimization is provided in Figure 5.9 featuring the hyperparameter importance and the hyperparameter relationships. Optuna uses a functional analysis of variance (fANOVA) (Hutter et al., 2014), to determine the hyperparameter with the strongest impact on the target metric. The search shows, that for the BERT models the learning rate has the strongest impact on the model performances. For the feature-based models, the "min_df" parameter has a strong importance for all models. Words that have a document frequency strictly lower than the threshold "min_df" are ignored in the TF-IDF analysis. For the MNB, the additive smoothing parameter mnb_alpha has the strongest importance.

Combining human- and machine-generated annotations

To evaluate to which extent the human- and machine-generated annotation could be improved, ensembling techniques over different groups are applied. An ensemble is created by averaging the confidence assignment of the annotators and the calibrated probabilities of the NLP models. In the human annotations, a confidence is only specified to the assigned label, therefore, a probability of zero was assigned to all remaining classes. The class with the highest probability after the ensembling process represents the final prediction. In case two classes have the same probability, the predicted class is randomly assigned to one of those. An ensembling is created across different subgroups of annotations, which are described in more detail in Section 5.3.2.

Evaluation and statistical tests

All evaluation results of the presented methods are presented in more detail in the next section. In summary, the TRC classification performance of the machine- and human-

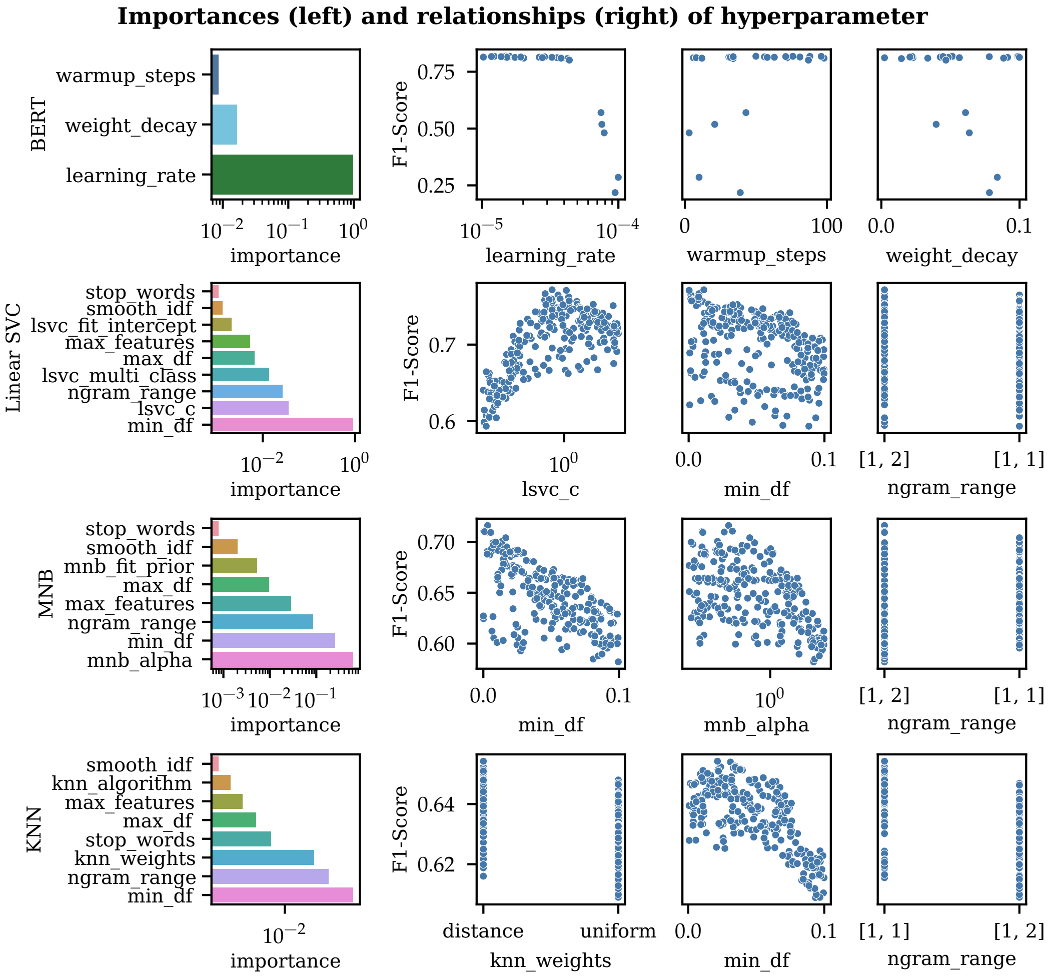


Figure 5.9: *Visualization of the hyperparameter importance and the weighted F1 score as a function of the hyperparameter as provided by Optuna (Akiba et al., 2019). The importance provided by Optuna using a functional analysis of variance. For the relationship plots, the three most decisive hyperparameters are depicted. For BERT, a low enough learning rate is crucial for a well-performing model. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, KNN=K-nearest neighbors. Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes.*

created baselines are reported using weighted recall, precision and F1 scores as well as accuracy. In addition, for the machine-created baselines, Area under the receiver operating characteristic curve (AUC) scores are calculated and receiver operating characteristic (ROC) curves are provided. For a rank-based evaluation and to assign confidence intervals to the different baselines, a bootstrap resampling with 2000 bootstrap samples is applied (Wiesenfarth et al., 2021; Efron et al., 1994; Efron, 2003). In detail, for each bootstrap sample, a ranking of the methods considered is created, which results in a rank distribution of the models' performances. To evaluate the calibration of the generated probabilities and to assess to which extent the confidence scores given by the human annotators are calibrated, the ECEs are reported and the accuracies for different confidence intervals, $M = 5$ bins, are plotted. Furthermore, correlations between different characteristics and metrics are qualitatively analyzed. Moreover, as described in Section 4.3.4 to qualitatively interpret the results, the attention weights, and the TF-IDF weights for a few sample reports are colorized. And for the BERT model, the CLS token representations which are the input for the classification layer, are visualized using a UMAP projection generated across all test sets. The embeddings are, respectively, colored with the TRC, token count and test set.

For all analyses and reported performance presented, different statistical tests are used to assess the significance of the results. Statistical tests include a t-test for continuous variables which follow a normal distribution, a MannWhitney U test for continuous variables which do not follow a normal distribution, a χ^2 test for categorical variables, and an Analysis Of Variance (ANOVA) for more than two categorical groups. Furthermore, a Tukey's honestly significant difference (Tukey's HSD) post hoc analysis is performed to analyze the variance between two groups.

5.3.2 Results

The report corpora

Table 5.9 gives an overview of the used datasets along with information on the characteristics of the patients. In total, 10455 reports form the final dataset with patients of an average age of 60±14 years including 5303 reports written for women and 5152 for men. The average number of visits per patient is around 2.2±1.6. In comparison to the training dataset SOR_{Train}, especially the two test datasets FTOR-DKFZ and FTORT-TKH differ in terms of the distribution of age, TRCs and tumor families as indicated by the P_{values}. This stems from the fact that each radiology department treats only patients in its field of oncologic expertise. Since the SOR_{Test} is sampled from the same pool of reports as the SOR_{Train}, there is no shift recognizable between the two datasets.

Besides the differences in the medical characteristics of the datasets, the distributional

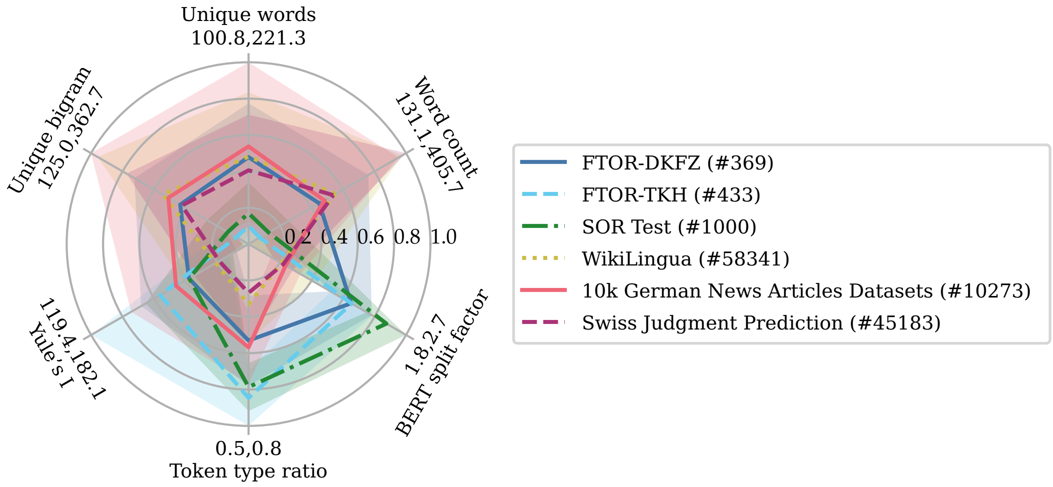


Figure 5.10: *Complexity measures of the three test datasets in comparison to those of three open-source German datasets. The shadows indicate the STD. Below the name of the characteristics, the range of the characteristics is given. For illustration purposes, the radial axis maps the ranges of the lexical complexity to the interval between zero and one. The number of samples per dataset is given in parenthesis behind the name of the dataset. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, SOR=Structured Oncology Report, STD=standard deviation.*

shift between the SORs and FTORs becomes apparent when examining the lexical complexity of the reports. In Table 5.10 the significance in differences is again indicated by the P_{values} in comparison to the SOR_{Train}. Therefore, the reports of the FTOR-DKFZ are much longer and richer in their complexity indicated by the high number of unique words and unique bigrams. The FTORT-TKH shows a significantly higher Yule's I metric than the other reports, which could be explained by the relatively short, but on-point reports. The higher token type ratio of the FTORT-TKH in comparison to the FTOR-DKFZ underlines this hypothesis. The introduced BERT split factor indicates that on average every word in the text needs to be split two and a half times till the word snippets are part of the vocabulary, determined during pretraining. Figure 5.10 illustrates the lexical complexity of the reports in comparison to the three public datasets WikiLingua (#58341), 10k German news articles (#10273) and Swiss Judgement Prediction (#45183) (cf. Table B.1 in the appendix for the tabular values). The lower BERT split factor shows that the vocabulary of the other datasets is better known to the pre-trained BERT model. The higher token type ratio and Yule's I metric

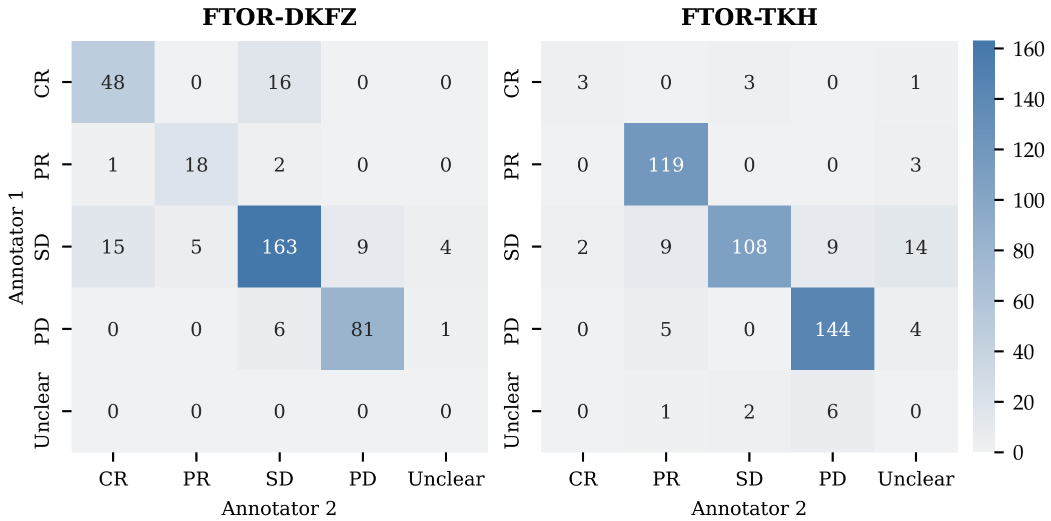
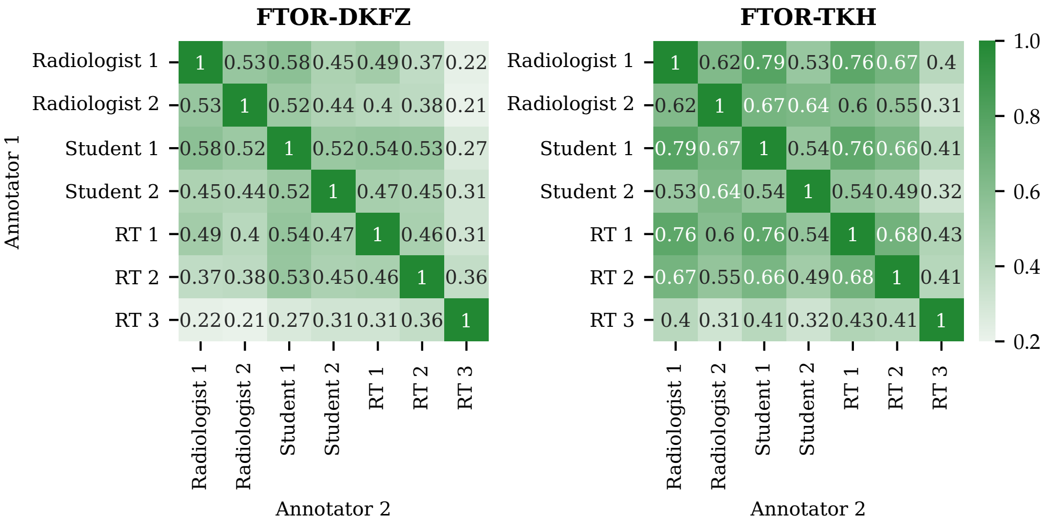


Figure 5.11: *Confusion matrices resulting from the export annotation of the FTOR-DKFZ and FTORT-TKH. The two unblinded radiologists (reference raters) assigned one of the four TRCs: complete response (CR), partial response (PR), stable disease (SD), progressive disease (PD) or an unclear label to a report based on the general information, findings and impression section of the FTORs. For all reports labeled with an unclear label and for reports of disagreement a subsequent consensus review was executed by the two readers to create a gold standard with the TRC annotation. This figure was adapted from Fink et al. (2022a). CR=complete response, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, PD=progressive disease, PR=partial response, SD=stable disease.*

demonstrate the high complexity of the medical reports.

Human annotations

The results of the reference annotations of the two radiologists for the FTOR-DKFZ and FTORT-TKH yielded a Cohen's kappa of 0.77 (0.71, 0.82) for the FTOR-DKFZ and of 0.90 (0.86, 0.93) for the FTORT-TKH. The confusion matrix in Figure 5.11 shows that most discrepancies occurred for labeling SD and CR (97 out of 118). For, in total, 118 of 802 cases of disagreement a consensus review was executed.
The results of the TRC classification of the seven blinded readers differ strongly between the two datasets FTOR-DKFZ and FTORT-TKH. Figure 5.13 illustrates in a heatmap the Cohen's kappa between the different annotators. The cohen's kappa scores are on average lower for the FTOR-DKFZ than for the FTORT-TKH dataset.



FTOR-DKFZ

Annotator 1 \ Annotator 2	Radiologist 1	Radiologist 2	Student 1	Student 2	RT 1	RT 2	RT 3
Radiologist 1	1	0.53	0.58	0.45	0.49	0.37	0.22
Radiologist 2	0.53	1	0.52	0.44	0.4	0.38	0.21
Student 1	0.58	0.52	1	0.52	0.54	0.53	0.27
Student 2	0.45	0.44	0.52	1	0.47	0.45	0.31
RT 1	0.49	0.4	0.54	0.47	1	0.46	0.31
RT 2	0.37	0.38	0.53	0.45	0.46	1	0.36
RT 3	0.22	0.21	0.27	0.31	0.31	0.36	1

FTOR-TKH

Annotator 1 \ Annotator 2	Radiologist 1	Radiologist 2	Student 1	Student 2	RT 1	RT 2	RT 3
Radiologist 1	1	0.62	0.79	0.53	0.76	0.67	0.4
Radiologist 2	0.62	1	0.67	0.64	0.6	0.55	0.31
Student 1	0.79	0.67	1	0.54	0.76	0.66	0.41
Student 2	0.53	0.64	0.54	1	0.54	0.49	0.32
RT 1	0.76	0.6	0.76	0.54	1	0.68	0.43
RT 2	0.67	0.55	0.66	0.49	0.68	1	0.41
RT 3	0.4	0.31	0.41	0.32	0.43	0.41	1

Figure 5.12: *Heatmap, reporting the inter-annotator agreement using the Cohen's kappa between the seven human annotators (two radiologists, two students and three RTs). The human annotators were only presented with the general information and the findings section. This figure was adapted from Fink et al. (2022a). FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, RT=radiology technologist.*

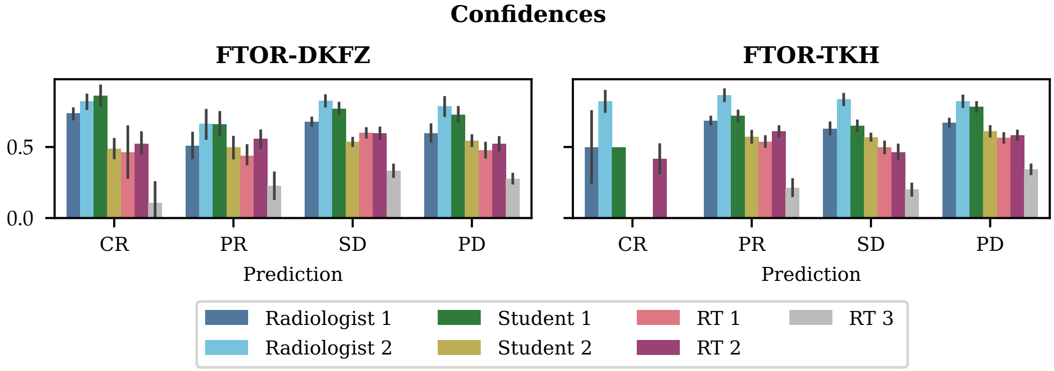


Figure 5.13: *Average confidence of the seven annotators on the FTOR-DKFZ and FTORT-TKH as a function of the TRCs present in the reports along with the STDs. CR=complete response, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, PD=progressive disease, PR=partial response, SD=stable disease, RT=radiology technologist, TRC=tumor response category.*

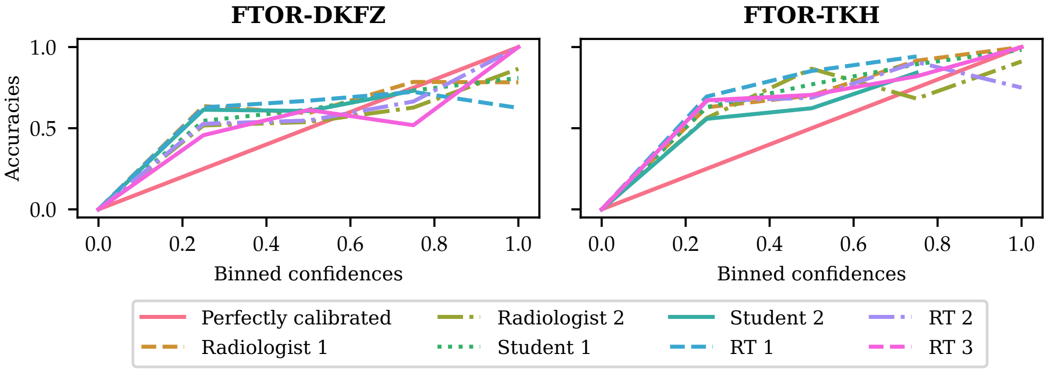


Figure 5.14: *Accuracies within binned confidence intervals of the seven annotators on the FTOR-DKFZ and FTORT-TKH along with a curve representing perfectly calibrated probabilities. FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, RT=radiology technologist.*

While the scores tend to be lower depending on the expertise level of the annotators, some outlier exists, e.g., between Radiologist 1 and Student 1 as well as Radiologist 1 and RT 1 for the FTORT-TKH. The confidences along with the performances of the three annotator groups are outlined in Table 5.12. The performance across multiple annotators or NLP models is calculated using micro-averaging, denoted as "MiA". In detail, the performance metric is calculated over a concatenated list of labels and predictions that belong to the annotations of the considered group. Table B.2 in the appendix, reports the confidences and performances of the single annotators. The confidences exhibit no significant differences between the two FTORs ($P_{value} = 0.26$), however. among the three annotator groups differences are significant ($P_{value} < 0.001$). Therefore, the radiologists have the highest confidences, followed by the medical students and the RTs. For a more fine-grained view, the confidences in dependence of the TRC are plotted in Figure 5.13. The figure indicates no clear trend of higher confidence for a certain TRC. Similarly to the confidences, the radiologists achieve the best F1 Score followed by the students and the RTs, however, the performances on the FTORT-TKH are better than on the FTOR-DKFZ for all annotator groups. To which extent the confidences assigned by the different annotators are calibrated on the two FTOR datasets is illustrated in Figure 5.14. The slightly S-shaped curves indicate not perfectly calibrated confidences. The ECE losses for all annotators, as well as machine-based methods, are reported together in Table 5.11.

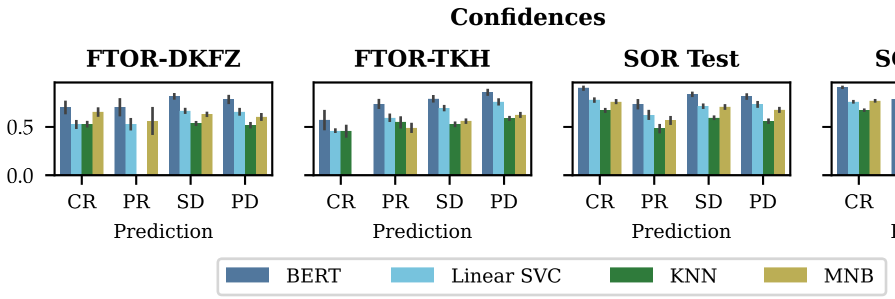


Figure 5.15: *Average confidence of the four NLP models on the the FTOR-DKFZ, the FTORT-TKH, the* $\mathrm{SOR}_{\mathrm{Test}}$ *and the* $\mathrm{SOR}_{\mathrm{Train}}$ *as a function of the TRCs present in the reports along with the STDs. BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing, PD=progressive disease, PR=partial response, SD=stable disease, SOR=Structured Oncology Report, TRC=tumor response category.*

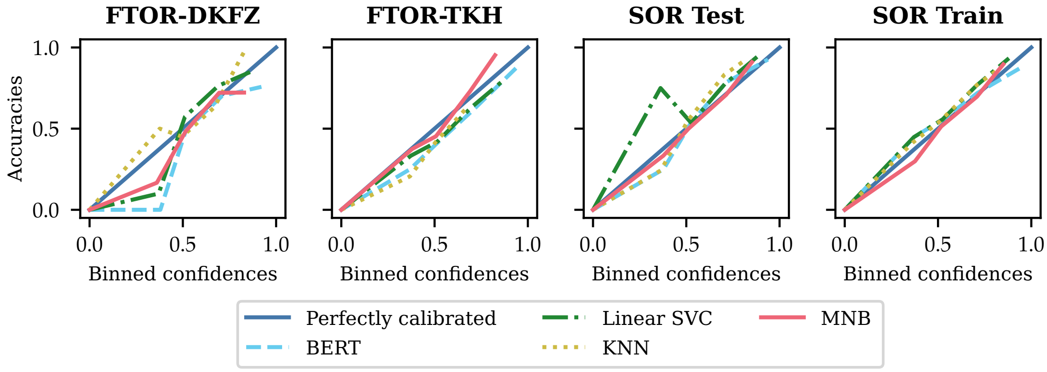


Figure 5.16: *Accuracies within binned confidence intervals of the four NLP models on the FTOR-DKFZ, the FTORT-TKH, the* $\mathrm{SOR}_{\mathrm{Test}}$ *and the* $\mathrm{SOR}_{\mathrm{Train}}$ *along with a curve representing perfectly calibrated probabilities. BERT=Bidirectional Encoder Representation from Transformers, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing, RT=radiology technologist, SOR=Structured Oncology Report, TRC=tumor response category.*

NLP models runs

In Table 5.13 the performances of the four NLP models on the SOR_{Test} and the two FTORs are reported. The BERT-based model outperforms the conventional NLP by a big margin on the SOR_{Test} and the FTORT-TKH, however, on the FTOR-DKFZ, the Linear-SVC can reach equal performance. The performance on the FTORs is for all algorithms significantly lower than on the SOR_{Test} underlining the difficulty of the distributional shift between the datasets. Similar to the human annotations, Figure 5.15 illustrates the confidences of the different NLP models in dependency of the TRC for the three test datasets. It is noted, that in this case the confidence is determined by the highest probability for the TRC. Only within a dataset correlation between the TRC and the confidence might be identifiable. Figure 5.11 illustrates the calibration of the models. The generated probabilities are relatively well calibrated with exceptions on the lower and upper probability borders, especially in the case of the FTOR-DKFZ. In Table 5.11 the corresponding ECE losses are reported.

Combining human- and machine-generated annotations

Figure 5.17 illustrates the performance differences of the plain human- and machine-based annotations in comparison to those calculated using micro-averaged or created using ensembling over different subgroups. An ensembling is created among the annotator groups, RTs, students and radiologists as well as across the two models BERT and Linear-SVC, since those perform best on the FTOR-DKFZ dataset. The combination of the BERT and Linear-SVC is denoted as "Machines". Furthermore, micro-averaged scores are calculated and ensembles are created between the different human annotations and the machines annotations to see how the machine-generated annotations could improve those of human annotators. It should be noted that in the case of micro-averaging, the F1 score is calculated over a concatenated set of predictions while when using ensembling, the F1 score is directly calculated on the merged predictions. For this reason, the comparison of the scores might be questionable. Figure 5.17 shows that ensembling surpasses the reported micro-averaged scores in most cases and that combining human annotations with machine-generated annotations, leads mostly to performance improvements for the FTOR-DKFZ, in contrast to the FTORT-TKH, for which performances losses are observed. A more detailed interpretation of the results is given in Section 6.3.5.

Evaluation and statistical tests

For a better comparison of the human and machine-based performances Figure 5.18 illustrates the ROC curve for the machine-based methods along with the operating points for the three annotator groups, respectively, for each TRC and a weighted

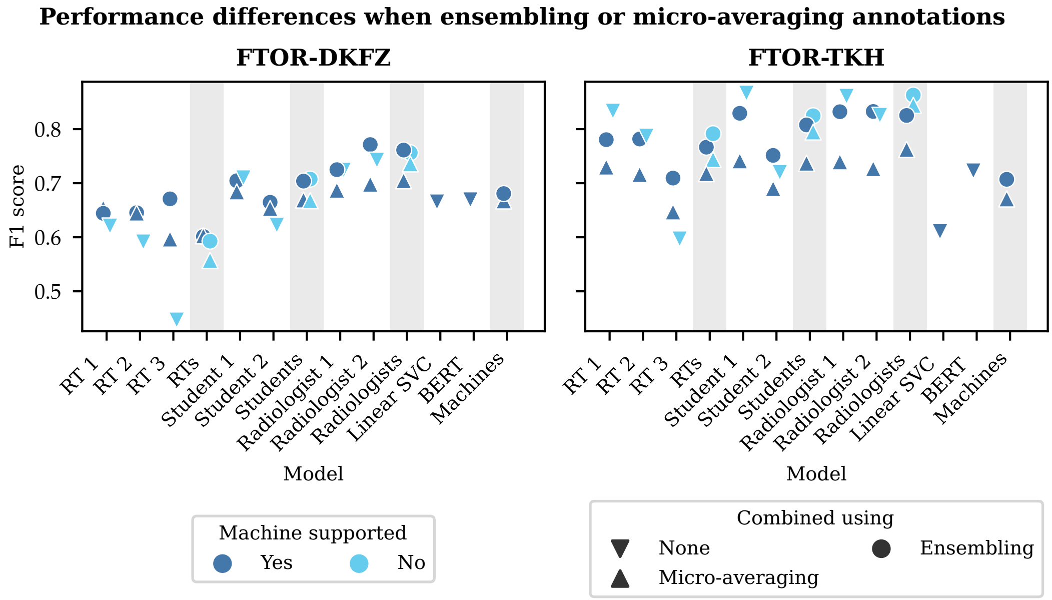


Figure 5.17: *Performance differences when ensembling or micro-averaging annotations over different subgroups. Drawn is the F1 score as a function of the annotation performances of the seven annotators, the Linear-SVC and the BERT model (white background, light blue markers), along with the micro-averaged performances and the ensembling performance across the groups RTs, students, radiologists (gray background, light blue markers). In the plot, the combination of Linear-SVC and the BERT model is denoted as "Machines". The performances are reported for micro-averaging or ensembling the two models (gray background, dark blue markers). In addition, all single and grouped human annotations are combined using micro-averaged or ensembling with the annotations of the "Machines" (white and gray background, dark blue markers). Whether the performances are calculated on the plain predictions, the ensembled predictions, or via micro-averaged is distinguished using different markers. BERT=Bidirectional Encoder Representation from Transformers, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, RT=radiology technologist.*

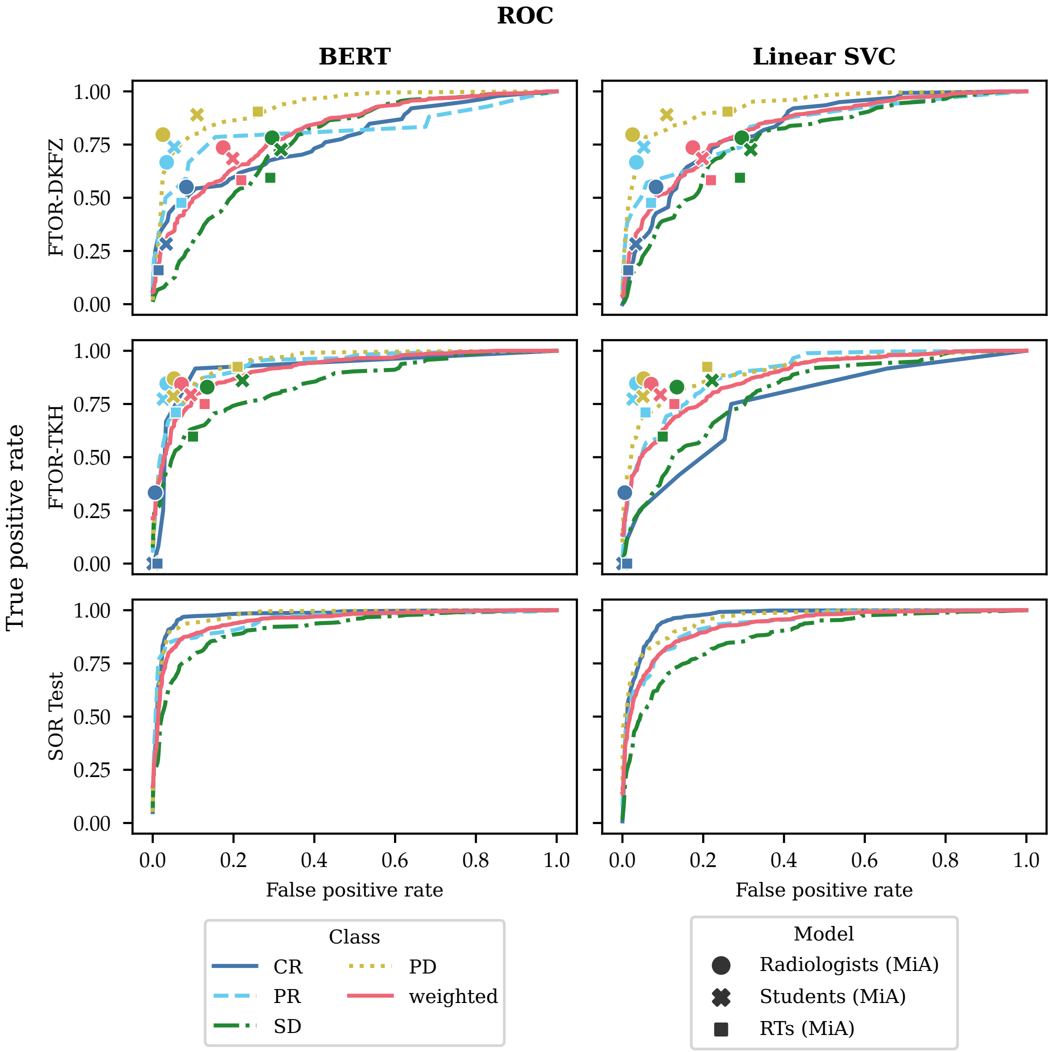


Figure 5.18: *ROC curves for the BERT model and the Linear-SVC on the three test datasets along with the operating points for the three annotator groups on the two FTORs. ROC curves are plotted, respectively, for each TRC along with a weighted average over all TRCs. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, FTOR=free-text-oncology report, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), ROC=receiver operating characteristic, RT=radiology technologist, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

average. The corresponding AUC scores are reported in Table 5.14. The scores and curves indicate that both, humans and machines, have difficulties with the SD class, whereas the PD class is one of the easiest.

Figure 5.19 illustrates the results of the ranking analysis over all annotators and machine-based approaches. It clearly shows that the machine-based models cannot reach the highest human baselines, given by the radiologists and in the case of the FTORT-TKH by Student 1. Whereas on the FTOR-DKFZ, both, the BERT-based model and the Linear-SVC outperform the RTs and Student 2, the models are distributed in the rank among annotators for the FTORT-TKH. The results emphasize again the almost equal performance of the BERT-based model and the Linear-SVC.

To illustrate a possible use case of the presented methods in clinical practice, timelines of the tumor-burdon change on a patient level are created and visualized in Figure 5.20. Along with the change in tumor-burdon of the patient over time, the confidence of the models at each time point is provided.

Correlations between the performances and properties of the datasets

Figure 5.21 illustrates the human- and machine-based performance in dependency of the grouped confidence for the averaged confidences of the annotators, the confidences of the BERT-based model and the confidences of the Linear-SVC. The figure shows that the performances of the machines and humans correlate with the confidences of the annotators as well as with the confidences generated by the BERT model. However, the machine's and human's performances do qualitatively not correlate with the confidences assigned by the Linear-SVC, suggesting that the yielded confidences of the BERT model are more human-like than those of the Linear-SVC.

The correlations between the performances of the human and machine annotations in dependency of grouped lexical complexity are illustrated in Figures 5.22 and 5.23. The figures suggest varying profound correlations between the F1 Scores and the considered measures. For the FTOR-DKFZ the performance of almost all annotations decreases with higher word counts, unique words and unique bigrams, only the radiologists are not influenced by the length of the reports. The effect is almost inverse (except for the RTs) for the much more concise, highly disease-specific FTORT-TKH, in which humans and machines profit from longer reports with higher numbers of unique words and unique bigrams. In addition reports with a high complexity indicated by high Yule's I and token type ratio for the FTORT-TKH a visible drop in performance is recognizable. The effect does not translate to the FTOR-DKFZ. The performance of BERT does not show a visible performance change with a varying BERT split factor.

A further interesting correlation is how human and machine confidences and performances are influenced in case of agreement and disagreement between oncologic and nononcologic findings. This analysis is only done on the FTORT-TKH. i.e., oncologic to nononcologic concordance (agreement) is annotated for the findings increased pulmonary metastases and increased degenerative changes of the spine in one FTOR, whereas a discordance (disagreement) is annotated for the findings "decreased pulmonary metastases and increased degenerative changes of the spine in another FTOR. Figure 5.24 shows that the confidences of the humans are not significantly different, in contrast to the machine-generated confidences which show a significant decrease ($P_{value} < 0.0001$ and $P_{value} < 0.001$) in confidence in case the oncologic and nononcologic findings disagree. Figure 5.25 illustrates the F1 Scores in dependency of the agreement and disagreement of oncologic and nononcologic findings for the machines and human annotations. Especially, the machine performances strongly decrease in case of disagreeing oncological and nononcologic findings.

Interpretability and explainability of the models

In terms of interpretability, the attention weights generated by the fine-tuned BERT model as well as the TF-IDF weights used for the Linear-SVC model are qualitatively plotted in Figure 5.26 for one sample of the FTORT-TKH. The weights of the five folds are averaged in the representation and respectively mapped to the interval between 0 and 1. Special tokens such as the CLS token or SEP token of the BERT model are removed prior to the mapping. In the example, all models predict the wrong TRC. However, words indicating a certain TRC such as "deutlich weniger (significantly less)" or "vergößerter lymphknoten (enlarged lymph node)" receive high weights for all models. The interpretability of the attention weights from BERT is still the subject of research. To only qualitative assess tendencies in the attentions weights Figure 5.27 illustrates the attention weights towards the CLS token token for the pre-trained model and, respectively, for the models of the five folds. The illustration shows that attention weights from the pre-trained model persist in the fine-tuned model in a weakened form. However, as mentioned above, tokens indicating a certain TRC receive stronger attention. Furthermore, Figure 5.28 visualizes the CLS token embeddings colorized, respectively, with the TRC, the token count and the test set. The separation of labels shows the qualitative learned TRC. The coloring of token counts shows that the vectors, average for sequences longer than 512 specific subclusters emerged within those the classification seems to be learned individually. The coloring of the test set shows that also after fine-tuning the sets differ in their vector representation, complicating the generalization of the learned model on unseen data.

Parameter	Explanation	BERT	Linear-SVC	KNN	MNB
learning_rate	Learning rate for the AdamW optimizer	0.000			
warmup_steps	Number of steps for the warmup phase	50.000			
weight_decay	Weight decay for the AdamW optimizer	0.099			
Max epochs	Maximal number of epochs for BERT. The best model of all epochs is taken for inference.	10.000			
C	Regularization parameter for the SVC		0.3572		
fit_intercept	Whether to calculate the intercept for the SVC model		True		
multi_class	Strategy for multi class, here, one-versus-rest		ovr		
algorithm	Algorithm to compute the nearest neighbors, here, brute-force search			brute	
weights	Weight function used for the prediction			distance	
alpha	Additive (Laplace/Lidstone) smoothing parameter				0.0293
fit_prior	Whether to learn class prior probabilities or not.				False
max_df	When building the vocabulary ignore terms that have a document frequency strictly higher than the given threshold			0.9619	0.9382
max_features	Build a vocabulary that only consider the top max_features ordered by term frequency across the corpus.			19664	5277
min_df	When building the vocabulary ignore terms that have a document frequency strictly lower than the given threshold			0.0215	0.0032
ngram_range	The lower and upper boundary of the range of n-values for different n-grams to be extracted.			(1,1)	(1,2)
smooth_idf	Smooth idf weights by adding one to document frequencies, as if an extra document was seen containing every term in the collection exactly once.			FALSE	TRUE
stop_words	Terms that are ignored for max_df, max-features and min_df			TRUE	FALSE

Table 5.8: *Best hyperparameter settings for the four NLP models after maximizing the weighted F1 score on the* n^{th} *validaiton fold using* $k = 5$ *folds. The random hyper-parameter search is executed using the software toolkit Optuna (Akiba et al., 2019) with 25 trials for BERT and, respectively, 250 trials for the other models. Explanations for the parameter names can be found in the documentation of scikit-learn (Pedregosa et al., 2011). BERT is trained using a batch size of 8, an Adam epsilon of 1e-8, and a maximum number of 10 epochs while using the best model out of all epochs for testing. This table was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing.*

Dataset	All (#10455)	All SORs (#9653)	SOR$_{Train}$ (#8653)	SOR$_{Test}$ (#1000)	FTOR-DKFZ (#369)	FTORT-TKH (#433)
Patients						
Age (y)*	61±14	60±14	60±14	60±14	65±15	65±9
P Value vs SOR$_{Train}$				0.42	<.001	<.001
Sex						
Women	5303 (50.7)	4939 (51.2)	4435 (51.3)	504 (50.4)	194 (52.6)	170 (39.3)
Men	5152 (49.3)	4714 (48.8)	4218 (48.7)	496 (49.6)	175 (47.4)	263 (60.7)
P Value vs SOR$_{Train}$				0.63	0.66	<.001
TRC						
PD	2467 (23.6)	2208 (22.9)	1979 (22.9)	229 (22.9)	91 (24.7)	168 (38.8)
SD	4018 (38.4)	3701 (38.3)	3318 (38.3)	383 (38.3)	188 (50.9)	129 (29.8)
PR	942 (9.0)	791 (8.2)	709 (8.2)	82 (8.2)	21 (5.7)	130 (30.0)
CR	3028 (29.0)	2953 (30.6)	2647 (30.6)	306 (30.6)	69 (18.7)	6 (1.4)
P Value vs SOR$_{Train}$				>.99	<.001	<.001
Tumor Families†						
Gastrointestinal	2423 (28.6)	2347 (30.8)	2115 (31.0)	232 (29.9)	28 (7.8)	48 (9.5)
Gynecologic	1868 (22.0)	1800 (23.7)	1625 (23.8)	175 (22.6)	62 (17.2)	6 (1.2)
Urogenital	1115 (13.2)	1075 (14.1)	969 (14.2)	106 (13.7)	25 (6.9)	15 (3.0)
Skin	873 (10.3)	781 (10.3)	701 (10.3)	80 (10.3)	92 (25.6)	
Lung	477 (5.6)	41 (0.5)	36 (0.5)	5 (0.6)	10 (2.8)	426 (84.4)
Soft tissue	415 (4.9)	409 (5.4)	370 (5.4)	39 (5.0)	6 (1.7)	
Head and neck	337 (4.0)	273 (3.6)	245 (3.6)	28 (3.6)	61 (16.9)	3 (0.6)
Liver	254 (3.0)	253 (3.3)	223 (3.3)	30 (3.9)	1 (0.3)	
Bone	226 (2.7)	225 (3.0)	199 (2.9)	26 (3.4)	1 (0.3)	
Biliary system	192 (2.3)	189 (2.5)	159 (2.3)	30 (3.9)	3 (0.8)	
CUP	177 (2.1)	161 (2.1)	143 (2.1)	18 (2.3)	16 (4.4)	
Lymphatic	45 (0.5)	14 (0.2)	12 (0.2)	2 (0.3)	24 (6.7)	7 (1.4)
Vascular	30 (0.4)	29 (0.4)	25 (0.4)	4 (0.5)	1 (0.3)	
Hematologic	27 (0.3)	6 (0.1)	6 (0.1)		21 (5.8)	
Brain	14 (0.2)	5 (0.1)	5 (0.1)		9 (2.5)	
P Value vs SOR$_{Train}$				0.455	<.001	<.001

Table 5.9: *Patient characteristics for the different datasets. Reported are frequencies with percentages in parentheses if not indicated differently. Reports of FTORT-TKH are from a hospital specializing in chest diseases. In addition, the* P *Values between the SOR$_{train}$ and the three test datasets are reported. This table was adapted from Fink et al. (2022a). CR=complete response, CUP=cancer of unknown primary, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, PD=progressive disease, PR=partial response, SD=stable disease, SOR=Structured Oncology Report, TRC=tumor response category.*

**Data are means ± SDs*

†*Since some patients have been diagnosed with tumors from multiple tumor families the reported number does not sum up to the total number of reports. The reported percentage is calculated with respect to the total number of identified tumors and as a function of the total number of reports.*

Dataset Parameters	All (#10455)	All SORs (#9653)	SOR$_{Train}$ (#8653)	SOR$_{Test}$ (#1000)	FTOR-DKFZ (#369)	FTORT-TKH (#433)
Word count	170.4±86.7	165.4±74.6	165.7±74.8	163.0±72.9	347.0±179.0	131.1±39.6
P Value vs SOR$_{Train}$				.27	<.001	<.001
Unique words	123.8±51.2	121.7±47.8	121.9±47.9	120.1±46.7	205.1±81.5	100.8±24.1
P Value vs SOR$_{Train}$				.26	<.001	<.001
Unique bigram	159.7±77.5	155.6±68.6	155.9±68.8	153.2±67.2	306.8±148.0	125.0±36.4
P Value vs SOR$_{Train}$				.25	<.001	<.001
Yule's I	153.5±48.0	152.3±46.0	152.4±46.0	151.3±46.4	151.6±50.9	182.1±73.7
P Value vs SOR$_{Train}$				.46	.73	<.001
Token type ratio	0.8±0.1	0.8±0.1	0.8±0.1	0.8±0.1	0.6±0.1	0.8±0.1
P Value vs SOR$_{Train}$				.55	<.001	<.001
BERT split factor	2.6±0.2	2.6±0.2	2.6±0.2	2.7±0.2	2.4±0.2	2.4±0.2
P Value vs SOR$_{Train}$				.07	<.001	<.001

Table 5.10: *Results of the lexical complexity analysis. Means of the parameters are reported along with the 95% CI in parentheses. Furthermore, the* P *Values between the SOR*$_{train}$ *and the three test datasets are reported. This table was adapted from Fink et al. (2022a). CI=Confidence Interval, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, SOR=Structured Oncology Report.*

Model	FTOR-DKFZ	FTORT-TKH	SOR$_{Test}$	SOR$_{Train}$
BERT	0.101	0.078	0.026	0.053
KNN	0.049	0.076	0.088	0.049
Linear-SVC	0.064	0.075	0.056	0.051
MNB	0.039	0.050	0.025	0.023
RT 1	0.168	0.297		
RT 2	0.096	0.190		
RT 3	0.116	0.183		
Radiologist 1	0.102	0.181		
Radiologist 2	0.131	0.118		
Student 1	0.148	0.154		
Student 2	0.112	0.130		

Table 5.11: *Expected Calibration Error (ECE) loss of the four NLP models and the seven annotators on the three test datasets and the SOR*$_{Train}$*. Since no human annotation is executed on the SORs, also no ECE loss is reported for the human annotators on the SOR data splits. BERT=Bidirectional Encoder Representation from Transformers, ECE=Expected Calibration Error, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing, SOR=Structured Oncology Report.*

Dataset	Annotators	Confidence	Recall (%)	Precision (%)	Accuracy (%)	F1 Score
FTOR-DKFZ	Radiologists (MiA)	3.92±1.02	73.5 (70.5 ,76.3)	74.2 (71.3 ,77.0)	73.5 (70.5 ,76.3)	0.74 (0.71 ,0.76)
	Students (MiA)	3.58±1.04	68.5 (65.7 ,71.3)	68.8 (65.6 ,71.9)	68.5 (65.7 ,71.3)	0.67 (0.64 ,0.70)
	RTs (MiA)	2.86±1.10	58.3 (55.8 ,60.7)	62.8 (59.4 ,66.1)	58.3 (55.8 ,60.7)	0.56 (0.53 ,0.58)
FTORT-TKH	Radiologists (MiA)	4.01±0.93	84.3 (82.2 ,86.3)	85.0 (83.1 ,86.8)	84.3 (82.2 ,86.3)	0.84 (0.82 ,0.86)
	Students (MiA)	3.61±0.84	79.3 (77.1 ,81.4)	81.6 (79.7 ,83.4)	79.3 (77.1 ,81.4)	0.79 (0.77 ,0.81)
	RTs (MiA)	2.85±1.06	75.0 (73.0 ,77.1)	75.3 (73.1 ,77.3)	75.0 (73.0 ,77.1)	0.74 (0.72 ,0.76)
FTORs (MiA)	Radiologists (MiA)	3.97±0.97	79.3 (77.5 ,81.0)	79.7 (78.0 ,81.4)	79.3 (77.5 ,81.0)	0.79 (0.78 ,0.81)
	Students (MiA)	3.60±0.94	74.3 (72.6 ,76.1)	74.5 (72.5 ,76.4)	74.3 (72.6 ,76.1)	0.73 (0.72 ,0.75)
	RTs (MiA)	2.86±1.08	67.3 (65.7 ,68.8)	66.9 (64.9 ,68.8)	67.3 (65.7 ,68.8)	0.65 (0.63 ,0.67)

Table 5.12: *TRC classification results, micro-averaged, for the three human annotator groups, respectively, on the FTOR-DKFZ and FTORT-TKH and across all FTORs. Reported scores are mean values with 95% CIs in parenthesis unless otherwise noted. The confidences assigned during annotations are reported as means ± STDs. This table was adapted from Fink et al. (2022a). CI=Confidence Interval, FTOR=free-text-oncology report, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), RT=radiology technologist, SOR=Structured Oncology Report, STD=standard deviation, TRC=tumor response category.*

Dataset	Model	Confidence	Recall (%)	Precision (%)	Accuracy (%)	F1 Score
FTOR-DKFZ	BERT	0.79±0.16	69.4 (66.1 ,72.9)	72.4 (68.8 ,76.1)	69.4 (66.1 ,72.9)	0.67 (0.63 ,0.71)
	Linear-SVC	0.65±0.14	69.1 (65.6 ,72.4)	67.5 (63.1 ,71.8)	69.1 (65.6 ,72.4)	0.67 (0.63 ,0.70)
	KNN	0.53±0.10	49.7 (45.5 ,53.7)	47.0 (43.0 ,51.1)	49.7 (45.5 ,53.7)	0.47 (0.43 ,0.51)
	MNB	0.63±0.12	61.2 (57.5 ,65.0)	63.2 (58.9 ,67.3)	61.2 (57.5 ,65.0)	0.59 (0.55 ,0.63)
FTORT-TKH	BERT	0.81±0.17	73.0 (69.7 ,76.2)	74.8 (71.7 ,77.9)	73.0 (69.7 ,76.2)	0.72 (0.69 ,0.76)
	Linear-SVC	0.71±0.15	63.1 (59.8 ,66.5)	73.5 (70.2 ,76.6)	63.1 (59.8 ,66.5)	0.61 (0.57 ,0.65)
	KNN	0.56±0.11	48.6 (45.5 ,52.0)	57.8 (51.1 ,63.8)	48.6 (45.5 ,52.0)	0.42 (0.39 ,0.46)
	MNB	0.59±0.12	57.8 (54.7 ,61.0)	65.6 (61.5 ,69.2)	57.8 (54.7 ,61.0)	0.53 (0.50 ,0.57)
SOR_{Test}	BERT	0.84±0.15	85.2 (83.3 ,86.9)	85.1 (83.2 ,86.9)	85.2 (83.3 ,86.9)	0.85 (0.83 ,0.87)
	Linear-SVC	0.73±0.15	78.9 (76.9 ,80.9)	79.0 (76.8 ,81.0)	78.9 (76.9 ,80.9)	0.79 (0.76 ,0.81)
	KNN	0.61±0.12	68.7 (66.5 ,70.8)	69.0 (66.4 ,71.4)	68.7 (66.5 ,70.8)	0.68 (0.65 ,0.70)
	MNB	0.71±0.14	72.1 (69.9 ,74.4)	71.7 (69.5 ,74.1)	72.1 (69.9 ,74.4)	0.72 (0.70 ,0.74)
FTORs (MiA)	BERT	0.80±0.17	71.3 (69.1 ,73.7)	73.6 (71.1 ,76.1)	71.3 (69.1 ,73.7)	0.70 (0.67 ,0.73)
	Linear-SVC	0.68±0.15	65.8 (63.5 ,68.2)	68.7 (65.7 ,71.5)	65.8 (63.5 ,68.2)	0.63 (0.61 ,0.66)
	KNN	0.55±0.11	49.1 (46.5 ,51.6)	55.3 (50.9 ,59.2)	49.1 (46.5 ,51.6)	0.46 (0.43 ,0.48)
	MNB	0.60±0.12	59.4 (56.9 ,61.8)	64.3 (61.3 ,67.1)	59.4 (56.9 ,61.8)	0.56 (0.54 ,0.59)
ALL (MiA)	BERT	0.82±0.16	79.0 (77.6 ,80.5)	79.6 (78.1 ,81.0)	79.0 (77.6 ,80.5)	0.79 (0.77 ,0.80)
	Linear-SVC	0.71±0.15	73.1 (71.6 ,74.6)	74.4 (72.8 ,76.1)	73.1 (71.6 ,74.6)	0.72 (0.71 ,0.74)
	KNN	0.58±0.12	59.9 (58.3 ,61.5)	61.6 (59.4 ,63.7)	59.9 (58.3 ,61.5)	0.58 (0.56 ,0.59)
	MNB	0.66±0.14	66.4 (64.8 ,68.0)	67.1 (65.2 ,68.9)	66.4 (64.8 ,68.0)	0.65 (0.64 ,0.67)

Table 5.13: *TRC classification results of the four NLP models, respectively, on the FTOR-DKFZ and FTORT-TKH, across all FTORs and on the* SOR_{Test}[2]*. Reported scores are mean values with 95% CIs in parenthesis unless otherwise noted. The reported confidences, representing the highest probability assigned to a TRC class, are reported as means ± STDs. This table was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, FTOR=free-text-oncology report, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing, PD=progressive disease, PR=partial response, SD=stable disease, SOR=Structured Oncology Report, TRC=tumor response category.*

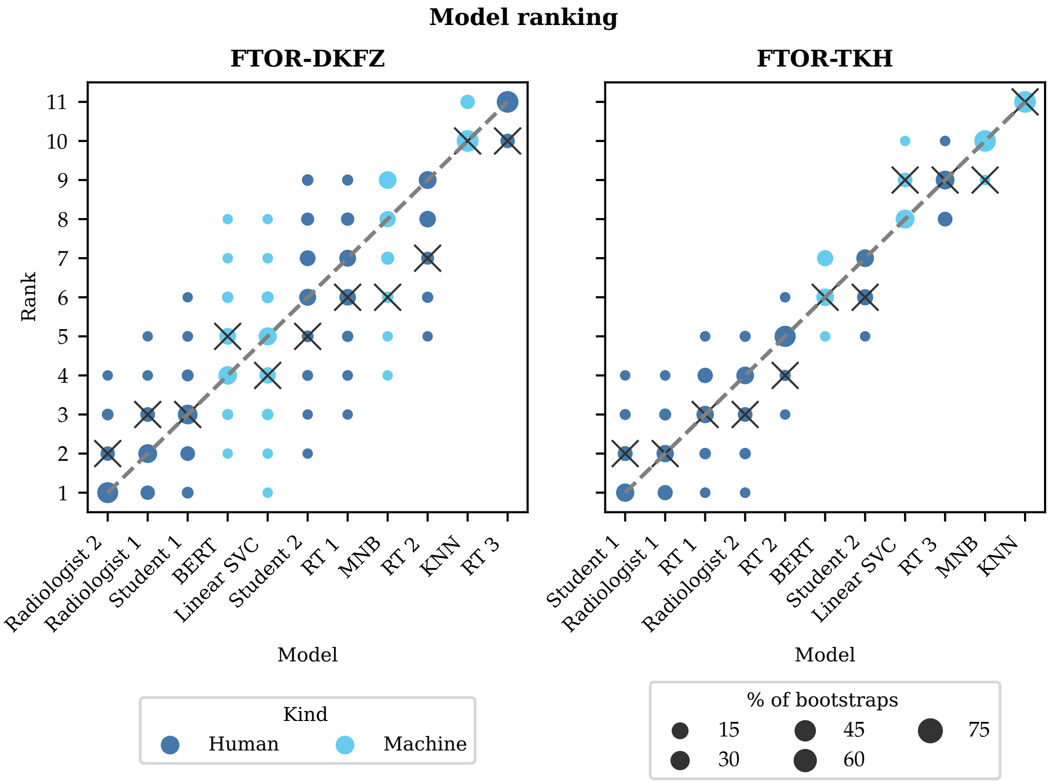


Figure 5.19: *Rank distributions of the seven human annotators and the four NLP models on the FTOR-DKFZ and FTORT-TKH. The methods are ranked respectively within each bootstrap sample with the best methods having the lowest rank. The human and machine models are differentiated by two different colors. The size of the points indicates the percentage of how often a method reached a rank among the 2000 bootstrap samples. The "X" indicates the median rank. of the methods. BERT=Bidirectional Encoder Representation from Transformers, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, NLP=Natural Language Processing, RT=radiology technologist.*

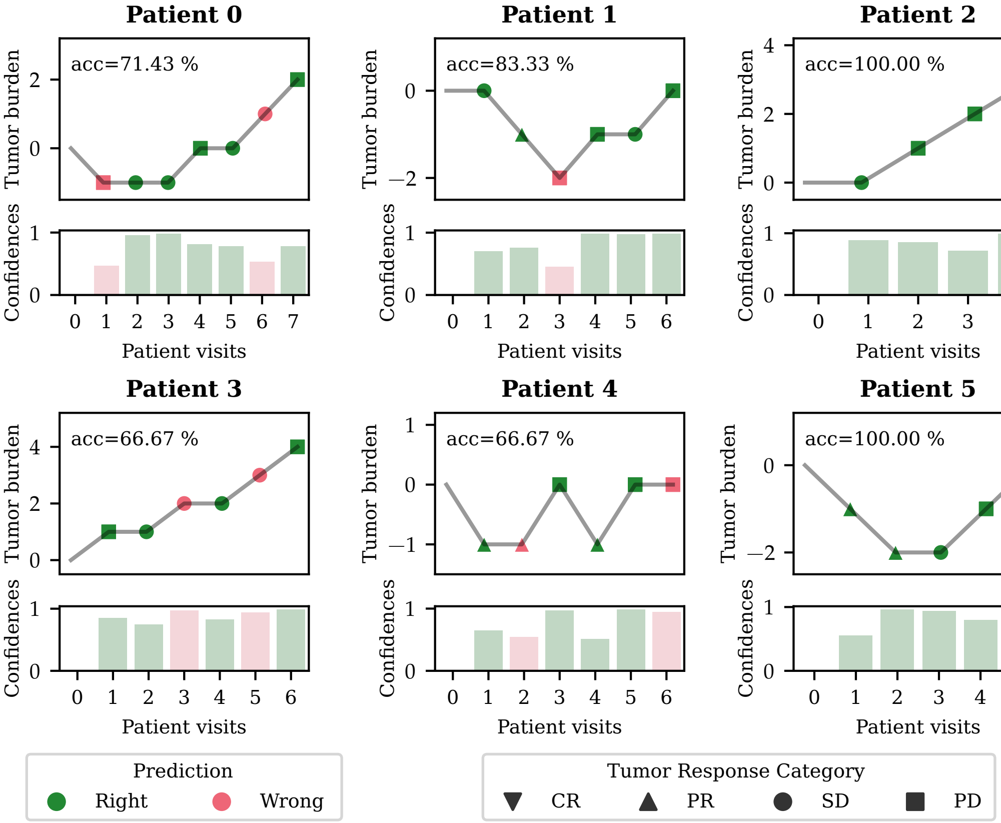


Figure 5.20: *Exemplary longitudinal representation of the oncologic course for six different patients using the TRCs, predicted by the BERT model. In addition, the confidence of the prediction, corresponding to the probability assigned to the predicted TRC, is reported using a bar plot. Wrong predictions are marked in red. Furthermore, the per-patient accuracy in the TRC classification is reported. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

Dataset	Model	CR	PD	PR	SD	weighted
FTOR-DKFZ	BERT	0.77 (0.77,0.78)	0.93 (0.93,0.93)	0.82 (0.82,0.82)	0.76 (0.76,0.76)	0.81 (0.81,0.81)
	Linear-SVC	0.83 (0.83,0.83)	0.93 (0.93,0.93)	0.84 (0.84,0.84)	0.78 (0.78,0.78)	0.83 (0.83,0.83)
FTORT-TKH	BERT	0.96 (0.96,0.96)	0.94 (0.94,0.94)	0.93 (0.93,0.93)	0.85 (0.85,0.85)	0.91 (0.91,0.91)
	Linear-SVC	0.77 (0.76,0.77)	0.91 (0.91,0.91)	0.90 (0.90,0.90)	0.80 (0.80,0.80)	0.87 (0.87,0.87)
$\mathrm{SOR}_{\mathrm{Test}}$	BERT	0.98 (0.98,0.98)	0.98 (0.97,0.98)	0.95 (0.95,0.95)	0.92 (0.92,0.92)	0.95 (0.95,0.95)
	Linear-SVC	0.97 (0.97,0.97)	0.96 (0.96,0.96)	0.94 (0.94,0.94)	0.88 (0.88,0.88)	0.93 (0.93,0.93)

Table 5.14: *AUCs scores of the BERT model and the Linear-SVC on the three test datasets, respectively, for each TRC along with the AUC score below the curve, weighted across all TRC. AUC=Area under the receiver operating characteristic curve, BERT=Bidirectional Encoder Representation from Transformers, CR=complete response, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

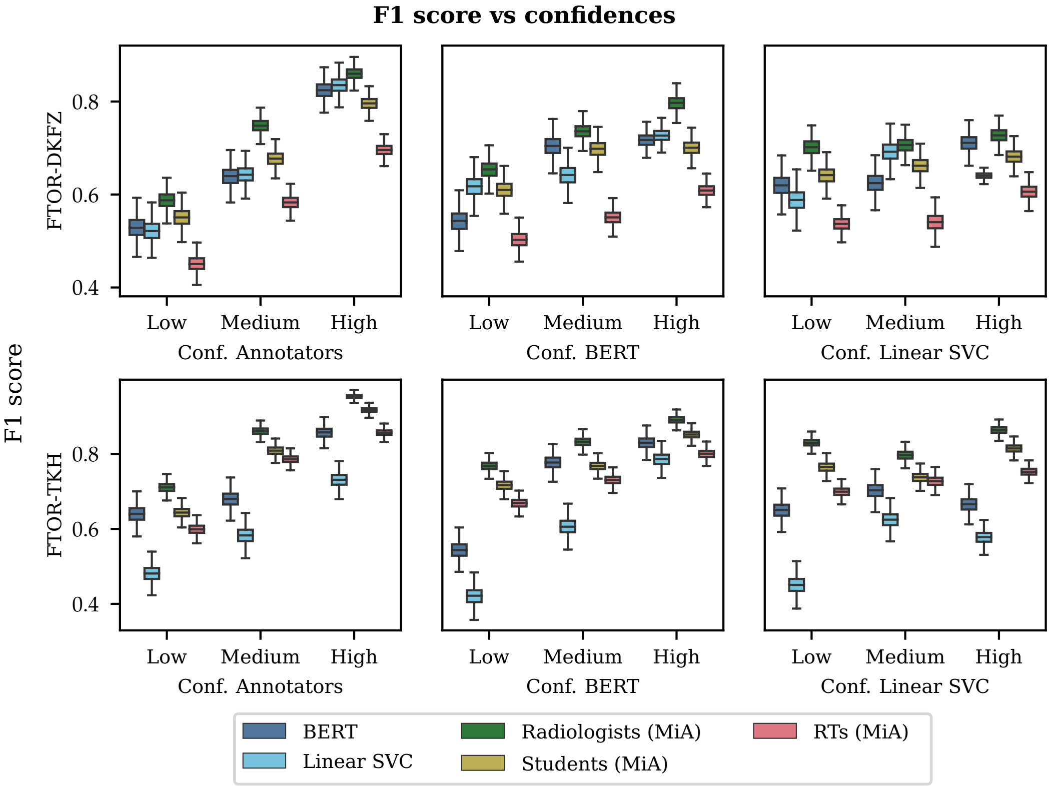


Figure 5.21: *Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of grouped confidence intervals of the annotators and two NLP models. For each subgroup, the lower and upper quartile, and the median position of the weighted F1 scores are illustrated. For the annotator groups, the performances of the individual annotators are aggregated using micro-averaging. The confidence groups are created by equal-sized bins of confidences. The annotator confidences are averaged across all the confidences assigned by the human annotators. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, NLP=Natural Language Processing, RT=radiology technologist.*

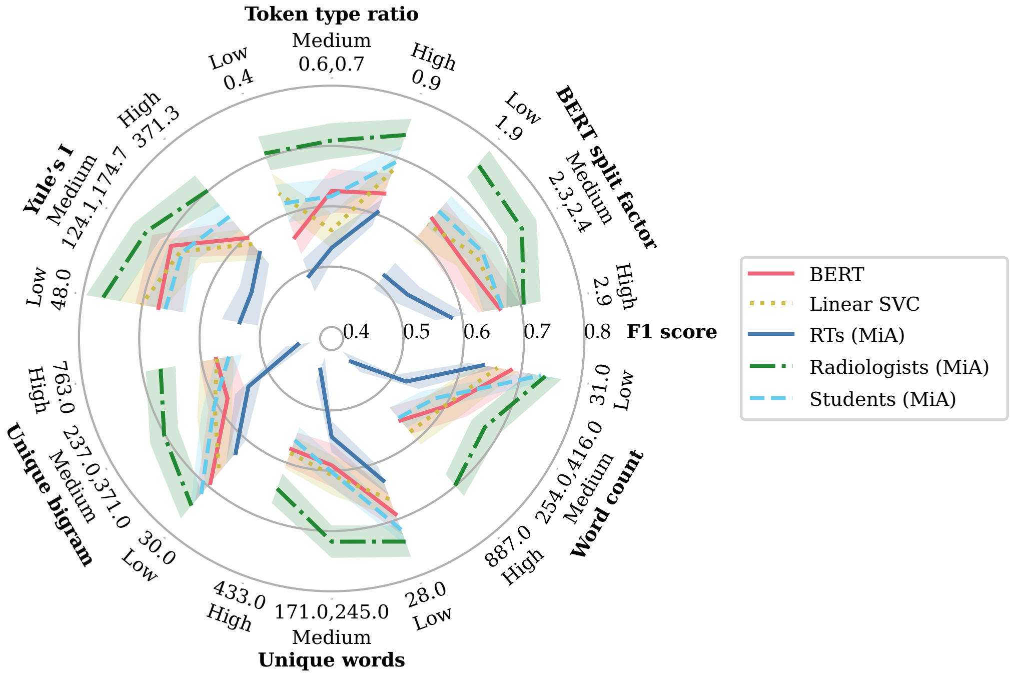


Figure 5.22: *Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the lexical complexity for the FTOR-DKFZ dataset. The three bins per complexity parameter are equally sized. The border of the bins is indicated below the name of the respective bin. The weighted F1 score is calculated within each bin and plotted on the radial axis. For the annotator groups, the performances of the individual annotators are aggregated using micro-averaging. Shadows indicate the 95% CI. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, CI=Confidence Interval, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, Linear-SVC=Linear Support Vector Classifier, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), RT=radiology technologist.*

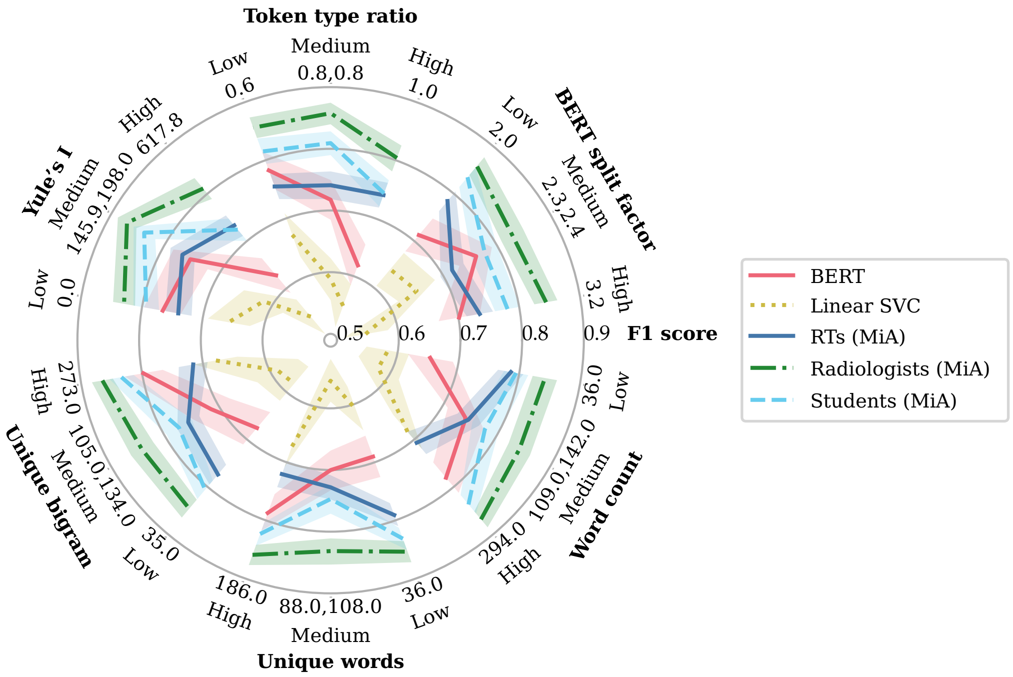


Figure 5.23: *Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the lexical complexity for the FTORT-TKH dataset. The three bins per complexity parameter are equally sized. The border of the bins is indicated below the name of the respective bin. The weighted F1 score is calculated within each bin and plotted on the radial axis. For the annotator groups, the performances of the individual annotators are aggregated using micro-averaging. Shadows indicate the 95% CI. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, CI=Confidence Interval, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), RT=radiology technologist.*

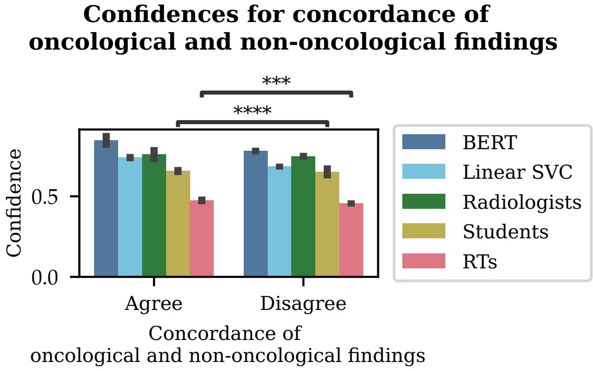


Figure 5.24: *Confidences of the BERT model, the Linear-SVC and the three annotators groups as a function of the concordance of oncologic and nononcologic findings described in the FTORT-TKH. For the three annotator groups, the reported confidences are averages. The difference in confidences for the radiologists is slightly significant (*$P_{value} < 0.0001$ *and* $P_{value} < 0.001$*), whereas, between the confidences of the annotators, no significant difference is present. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, RT=radiology technologist.*

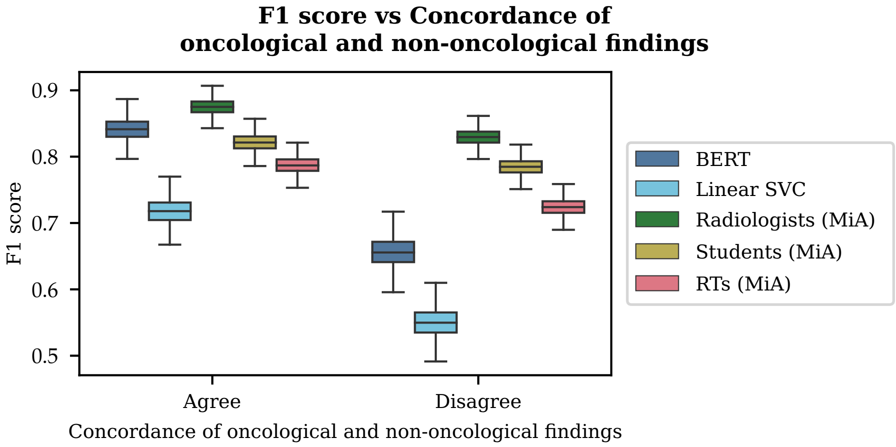
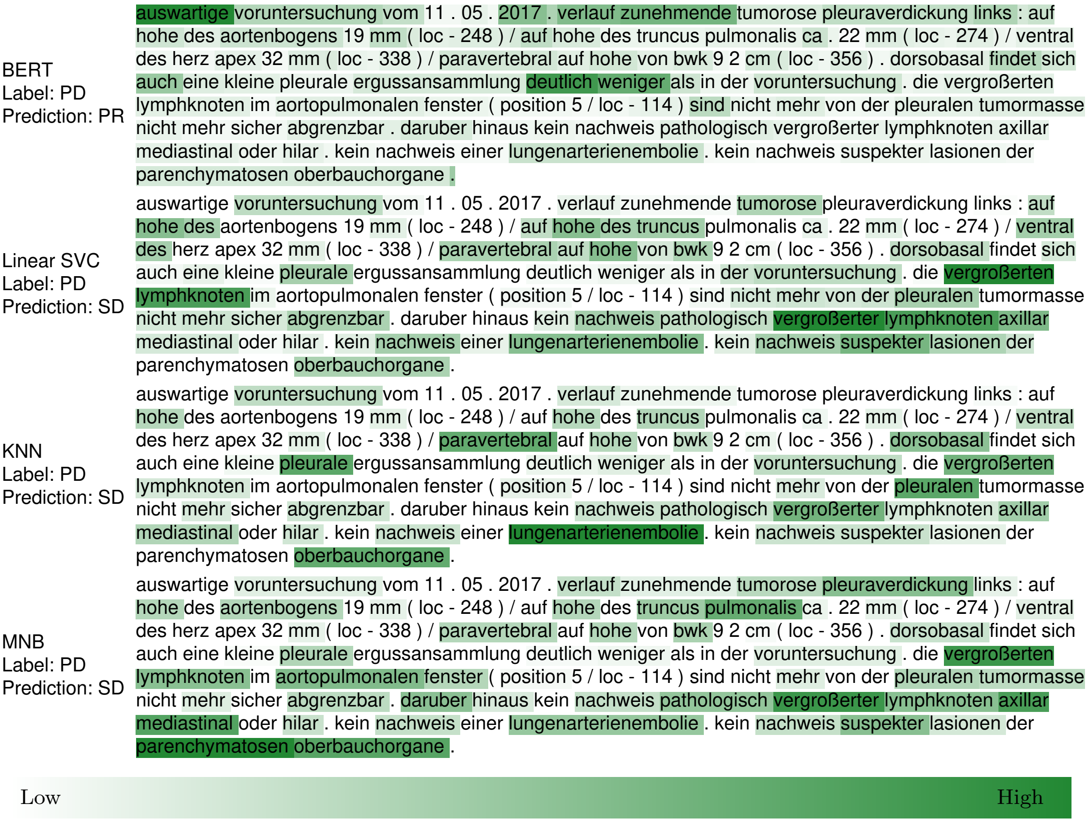


Figure 5.25: *Weighted F1 scores of the BERT model, the Linear-SVC and the three annotators groups as a function of the concordance of oncologic and nononcologic findings described in the FTORT-TKH. For each subgroup, the lower and upper quartile, and the median position of the weighted F1 scores are illustrated. For the annotator groups, the performances of the individual annotators are aggregated using micro-averaging. This figure was adapted from Fink et al. (2022a). BERT=Bidirectional Encoder Representation from Transformers, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, Linear-SVC=Linear Support Vector Classifier, RT=radiology technologist.*



BERT
Label: PD
Prediction: PR

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Linear SVC
Label: PD
Prediction: SD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

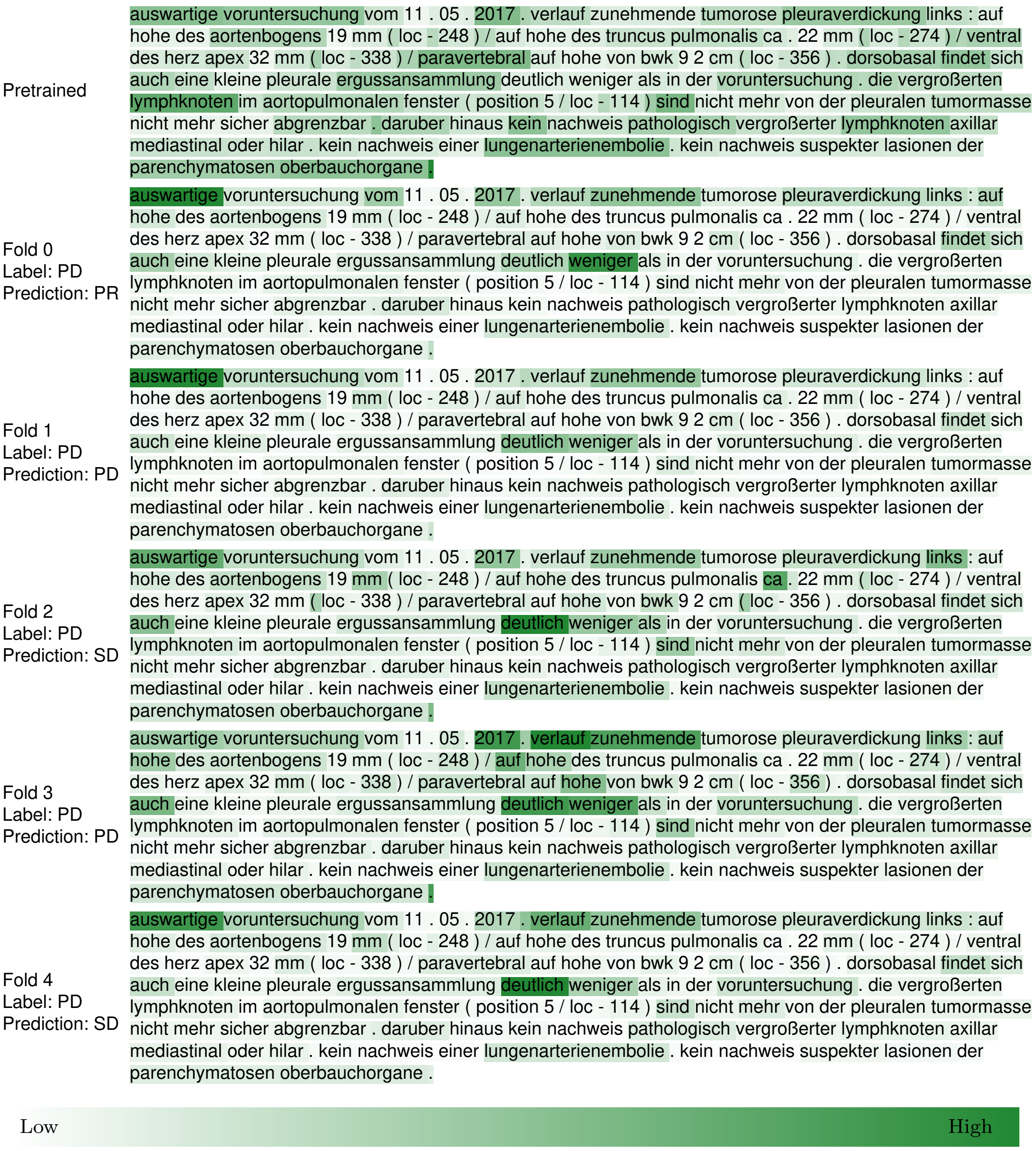
KNN
Label: PD
Prediction: SD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

MNB
Label: PD
Prediction: SD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Figure 5.26: *Visualization of token importance on a sample findings section of a FTORT-TKH for the four NLP models. For the three TF-IDF-based models the TF-IDF weights of the tokens are colorized in green. The stronger the tokens are colored, the stronger their importance. Tokens with a white background are not part of the maximum number of tokens considered in the TF-IDF analysis. For the BERT model the attention weights, averaged over all layers and heads are visualized. The minimum and maximum token importance is mapped into the interval between zero and one per finding. On the left, the name of the model and the reference label along with the predicted label are reported. BERT=Bidirectional Encoder Representation from Transformers, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, KNN=K-nearest neighbors, Linear-SVC=Linear Support Vector Classifier, MNB=Multinomial Naïve Bayes, NLP=Natural Language Processing, PD=progressive disease, PR=partial response, SD=stable disease, TF-IDF=term frequency-inverse document frequency.*



Pretrained

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Fold 0
Label: PD
Prediction: PR

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Fold 1
Label: PD
Prediction: PD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Fold 2
Label: PD
Prediction: SD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Fold 3
Label: PD
Prediction: PD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Fold 4
Label: PD
Prediction: SD

auswartige voruntersuchung vom 11 . 05 . 2017 . verlauf zunehmende tumorose pleuraverdickung links : auf hohe des aortenbogens 19 mm (loc - 248) / auf hohe des truncus pulmonalis ca . 22 mm (loc - 274) / ventral des herz apex 32 mm (loc - 338) / paravertebral auf hohe von bwk 9 2 cm (loc - 356) . dorsobasal findet sich auch eine kleine pleurale ergussansammlung deutlich weniger als in der voruntersuchung . die vergroßerten lymphknoten im aortopulmonalen fenster (position 5 / loc - 114) sind nicht mehr von der pleuralen tumormasse nicht mehr sicher abgrenzbar . daruber hinaus kein nachweis pathologisch vergroßerter lymphknoten axillar mediastinal oder hilar . kein nachweis einer lungenarterienembolie . kein nachweis suspekter lasionen der parenchymatosen oberbauchorgane .

Figure 5.27: *Visualization of token importance on a sample findings section of a FTORT-TKH for the BERT models trained on the* $k = 5$ *training folds along with the token importance of a completely untrained model. For the BERT model the attention weights, averaged over all layers and heads are visualized. The minimum and maximum token importance is mapped into the interval between zero and one per finding. On the left, the name of the model and the reference label along with the predicted label are reported. The predicted TRC varied for the generated folds. BERT=Bidirectional Encoder Representation from Transformers, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, PD=progressive disease, PR=partial response, SD=stable disease, TRC=tumor response category.*

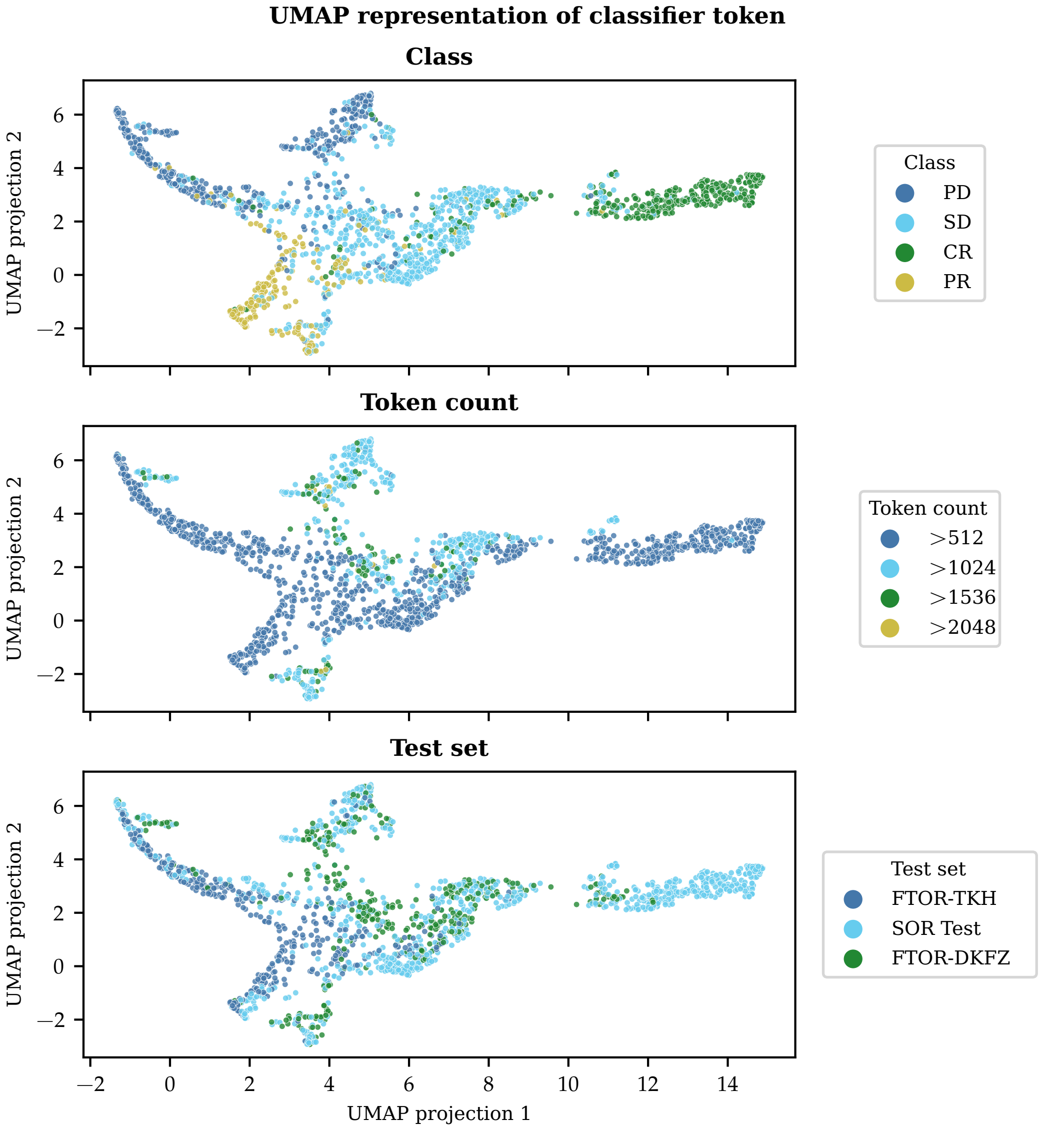


Figure 5.28: *Visualizations of the CLS token embeddings using UMAP projections on the three test datasets. The embeddings are colorized respectively with their TRC, their token count, and their corresponding test set. CLS token=Classifier token, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, PD=progressive disease, PR=partial response, SD=stable disease, SOR=Structured Oncology Report, TRC=tumor response category, UMAP=Uniform Manifold Approximation and Projection.*

5.4 Manual data annotation and tagging for medical images

5.4.1 Experiment setup

The implementation of the manual data annotation is evaluated based on four cases which are identified based on projects that use the Kaapana toolkit such as the JIP and RACOON. All use cases are executed on a locally running Kaapana platform. It is thereby ensured that the desired applications are technically feasible within the projects JIP and RACOON.

Use case 1 – Curating a cohort selection which was created based on DICOM metadata: A basic requirement for the development of good performing machine learning models is the quality of the training and validation data. Due to the huge variety of medical images inside clinical centers, a DICOM metadata query alone often includes samples that are not suitable for the application of medical image analysis algorithms. This use case enables the removal of those samples.

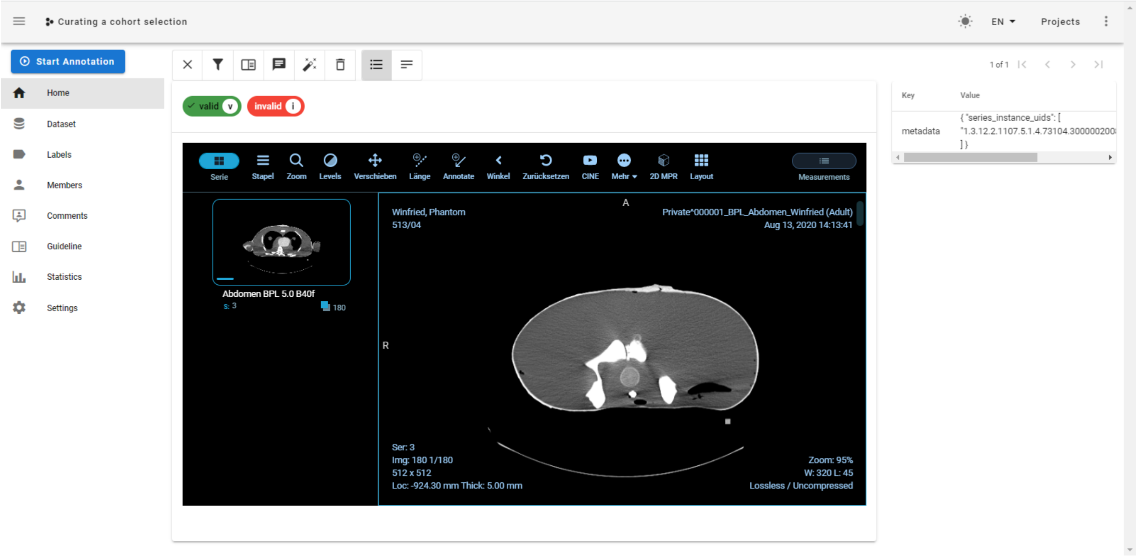
Use case 2 – Correcting falsy DICOM metadata: As pointed out in the introduction, the DICOM standard helps a lot when working with medical image data. However, despite the standard, not all metadata are always reliable. For this reason, a manual correction of that metadata is needed, which is evaluated in this use case.

Use case 3 – Adding free text to a DICOM image: Some medical imaging applications require further input in the form of free text such as image-to-text generation or text-to-image grounding.
The presented implementation for manual data annotation and tagging is part of the open-source project Kaapana (Scherer et al., 2023).

5.4.2 Results

In the following, the detailed steps of the different use cases are described. The use case descriptions are copied verbatim from Kades et al. (2022a) and in some cases slightly adapted.

Use case 1 – Curating a cohort selection which was created based on DICOM metadata: In this scenario, the platform user uses Kaapana's meta-dashboard to pre-define the desired cohort based on DICOM metadata. When sending the data to Doccano, they are prompted with an input screen where they can indicate that they want to create a classification project. In the next step, the user switches to Doccano,



Key	Value
metadata	{ "series_instance_uids": ["1.3.12.2.1107.5.1.4.73104.300000200] }

Figure 5.29: *Screenshot of the tagging process within Kaapana. The screenshot shows how the OHIF viewer is included in the Doccano classification view, which itself is embedded on the landing page of Kaapana. Instead of a text sample, the OHIF view of DICOM image corresponding to a specific study is placed in the middle of the screen. The assigned tags are visible at the top. This figure was adapted from Kades et al. (2022a) with permission of Springer. DICOM=Digital Imaging and Communications in Medicine, Kaapana=Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis, OHIF=Open Health Imaging Foundation.*

creates the necessary labels, e.g. a valid and an invalid image label, and assigns the labels to the corresponding images. Thanks to the keyboard shortcuts, only two clicks are required per image. After reviewing all images, the user can send the results back to Airflow, where the tags are added to the DICOM images. Figure 5.29 shows a screenshot of how tagging an image within the platform looks like (Kades et al., 2022a).

Use case 2 – Correcting falsy DICOM metadata: As in use case 1, the user can use the Kibana meta-dashboard to select images to edit and send the data to Doccano, specifying which DICOM tags to edit. A sequence-to-sequence project is created in Doccano. In the upper area, the user sees the image, and in the lower area, the specified DICOM tags that they can edit. Thanks to Doccano's user-friendly implementation, these steps can be performed with just a few clicks per sample. After completing the corrections on all DICOM images, they can save the corrections by triggering a workflow in Airflow that changes the DICOM metadata as specified in Doccano (Kades et al., 2022a).

Use case 3 – Adding free text to a DICOM image: Also in this use case, the user can define their cohort on the meta dashboard and send the data to Doccano. As in use case 2, a sequence-to-sequence project is created. The user can now add multiple diagnoses to the DICOM image. As before, the number of clicks is negligible compared to the time it takes to write the diagnosis. When finished, they can send the data back to Airflow where the diagnosis will be written to an appropriate DICOM tag (Kades et al., 2022a).

5.5 Real-world federated learning for the task of image segmentation

5.5.1 Experiment setup

The implementation presented in Section 4.5 to enable real-world federated learning for the nnU-Net is evaluated on the multi-site prostate MRI segmentation dataset. The implementation includes adjustments to the backend, the frontend, the local workflow and the nnU-Net itself. Besides the creation of a baseline by applying the implementation on the multi-site dataset, the results also demonstrate the feasibility of training the nnU-Net in a federated way, using the building blocks presented in Section 4.5. As an additional reference, the locally running implementation of the nnU-Net workflow in Kaapana is introduced in Section 3.1.2.

The multi-site prostate MRI segmentation dataset, introduced in detail in Section 3.2.3, consists of data from six different institutions with varying numbers of cases and acquisition protocols (Kades et al., 2022b; Liu et al., 2020a,b). In a preprocessing step, similar to Liu et al. (2020a), Liu et al. (2021) and Gu et al. (2021a), the PZ and CG segmentations of the RUNMC (Site A) and BMC (Site B) datasets are merged to work with a consistent reference definition across all sites (Kades et al., 2022b).

The experiments

Two different kinds of experiments are executed to benchmark the performance of differently trained nnU-Nets. All experiments are executed on six independent Kaapana instances which serve as clients during training. The federation and execution of the training are coordinated by an additional central Kaapana instance.

In the first experiment ("seen" setup), the datasets on each site are split into a training (70%) and test set (30%), respectively. The scenario is motivated by projects like RACOON featuring a high heterogeneity of image data, in which one aims to train an accurate model that performs equally well on all sites including the site(s) on which it was trained on. Table B.3 in the Appendix B.2 lists the identifier of the cases of each site for reproducibility and benchmarking. It has to be noted that in contrast to Jiang et al. (2022), the datasets are split on case- instead of slice-level. Moreover, in contrast to Gu et al. (2021a), no validation split is used (Kades et al., 2022b).

The idea behind the second experiment ("unseen" setup) is to evaluate the domain generalizability and robustness of the trained models toward sites that do not participate in the training. In those experiments, the leave-one-domain-out strategy applied by Liu et al. (2020a, 2021); Gu et al. (2021a) is adopted, in which the training is executed on K-1 seen sites and tested on the complete dataset of the omitted unseen target site. In contrast to Gu et al. (2021a), all cases of the unseen site are used for testing and no extra validation split of the unseen site is used to optimize the model.

Training of the nnU-Net

In all experiments, the nnU-Net is trained using different methods. The upper baseline method is a centralized training (denoted as "DeepAll") in which the nnU-Net is trained on all source domains combined. Additional baselines are created by training nnU-Nets on every single site independently, denoted as "Intra-site", as well as by ensembling the softmax predictions of the single-site trained models during testing, denoted as "Ensemble". For the leave-one-out experiments, ensembling is only applied to the models of the source domains. The final method is the federated training of the nnU-Net, denotes as "Federated". Existing works (Liu et al., 2020a; Jiang et al., 2022; Gu et al., 2021a) only trained 2D segmentation models due to the large variance in slice thickness between the different sites. In this work, 2D- and 3D-full-resolution model architectures are evaluated. All training runs are executed neither with a validation set nor with cross-validation, to keep the overall computational costs low and to train with as many cases as possible. By running preliminary single-site experiments, it is ensured that the models do not over- or underfit during training. All models are trained for 500 epochs with respectively 250 batches per epoch. The trained model is then used for testing. We did not tune any nnunet-specific hyperparameters, since those are either hard-coded or dynamically determined by the nnU-Net based on heuristics rules and the "dataset fingerprint". More information on how the hyperparameters are configured in the nnU-Net are given in Isensee et al. (2021).

For the federated training of the nnU-Net the $FedAvg$ algorithm (Brendan McMahan et al., 2016) is adopted and model updates are averaged after each epoch:

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{w_k^{(t+1)}}{K} \quad , \tag{5.2}$$

with w the parameters of the model, t the timestep, and the sum going over the number of participating clients K. To avoid biasing the optimization towards a particular site during training, the contributions of each client are weighted equally. This is done similarly in the ensembling method, but differently for the centralized method in which batches are sampled randomly from all available cases during training without considering the partitioning into sites. During training, a sum of cross-entropy and DICE loss is optimized. During testing, the performance is measured using the DICE score and the ASD. For reproducibility and the application of the presented methods, all implementations are available in the open-source project Kaapana (Scherer et al., 2023).

Table 5.15: *DICE (%) and ASD (mm) scores for all experiments, along with scores reported in works for existing methods. The last two columns report the scores and ranks, when averaged over the sites. Below the name of the dataset, the total number of cases per dataset is reported. Note that for the "seen" experiments, the test set includes 30% of all cases, whereas, in the "unseen" experiments, all cases are used for testing. The best scores for nnU-Net are marked as bold, and the best overall scores are underlined. This table was adapted from Kades et al. (2022b) with permission of Springer. DICE=Sørensen-Dice coefficient, nnU-Net=no new net U-Net.*

Setup	Dataset Algorithm	RUNMC # 30	BMC # 30	I2CVB # 19	UCL # 13	BIDMC # 12	HK # 12	Average	Rank
Seen	DCA-Net (Gu et al., 2021b)	91.83 0.72	91.59 0.81	89.93 0.77	91.99 0.64	90.68 0.93	90.57 0.82	90.93 0.78	
2D	Intra-site	87.74 0.79	91.14 0.72	81.12 2.05	88.06 0.82	69.83 2.35	85.11 1.08	83.83 1.30	5.05
	DeepAll	88.55 0.73	91.04 0.73	79.21 2.32	90.14 0.67	80.98 1.58	89.46 0.71	86.57 1.12	3.67
	Federated	88.27 0.77	90.88 0.70	**84.50** 2.00	**90.59** **0.61**	78.01 1.62	88.97 0.77	86.87 1.08	3.72
3D	Ensemble	87.48 0.92	86.27 3.46	48.28 20.93	88.02 0.88	58.32 15.54	82.51 8.00	75.15 8.29	6.12
	Intra-site	89.58 0.78	90.46 0.74	83.64 2.14	88.19 1.25	73.95 40.65	84.96 1.01	85.13 7.76	4.07
	DeepAll	**90.00** **0.67**	**91.57** 0.64	82.27 2.14	90.02 0.70	87.64 **1.26**	90.49 0.66	88.66 1.01	2.78
	Federated	89.96 0.69	91.50 **0.61**	**84.50** **1.95**	90.16 0.63	**87.70** 1.28	**90.99** **0.62**	**89.14** **0.96**	**2.60**
Unseen	SAML (Liu et al., 2020a)	89.66 1.38	87.53 1.46	84.43 2.07	88.67 1.56	87.37 1.77	88.34 1.22	87.67 1.58	
	ELCFS (Liu et al., 2021)	90.19	87.17	85.26	88.23	83.02	90.47	87.39	
	DCA-Net (Gu et al., 2021b)	90.61 1.12	88.31 1.14	84.89 1.76	89.22 1.09	86.78 1.58	89.17 1.02	88.16 1.29	
2D	DeepAll	84.89 1.37	83.10 1.26	71.17 4.54	85.88 **1.04**	74.18 4.73	86.24 1.20	80.91 2.36	3.22
	Federated	**85.84** **1.11**	81.96 **1.33**	**76.52** **4.52**	84.94 1.53	73.19 **2.56**	86.09 **1.03**	81.42 **2.01**	3.18
3D	Ensemble	76.53 38.57	84.99 2.25	49.14 37.49	84.34 16.68	72.15 18.96	85.81 5.72	75.49 19.95	3.56
	DeepAll	83.97 4.91	80.37 16.77	58.45 24.77	85.59 8.34	78.98 25.48	**89.24** 1.47	79.43 13.62	2.78
	Federated	85.01 3.65	**85.36** 8.05	67.63 16.34	**86.97** 1.78	**81.95** 21.16	88.51 1.86	**82.57** 8.81	**2.25**

5.5.2 Results

Feasibility of training the nnU-Net in a federated way

Figure 5.30 illustrates how the nnU-Net can be trained federated over multiple epochs using the building blocks that were added to Kaapana. The figure describes all steps as well as client-central communication during training in detail. Before any communication between the central instance and the client instances is possible, the instances must be connected by registering each other with individual authentication tokens. For the federated training, at the central instance, a job is submitted which triggers the nnU-Net federated DAG. Within the DAG, jobs are submitted to be executed on the participating client sites (1./7....). The client sites ask the central site periodically if a new job is available and fetches any new jobs to their site (3./9....), which triggers then locally the nnU-Net training DAG. This DAG contains in principle preprocessing steps, an operator for the actual training as well as postprocessing steps. In the initial preparation round, only the preprocessing of the nnU-Net training DAG is executed, which generates a fingerprint of the local datasets. This fingerprint is then shared with the central instance using a post-hook at the operator, which uploads a

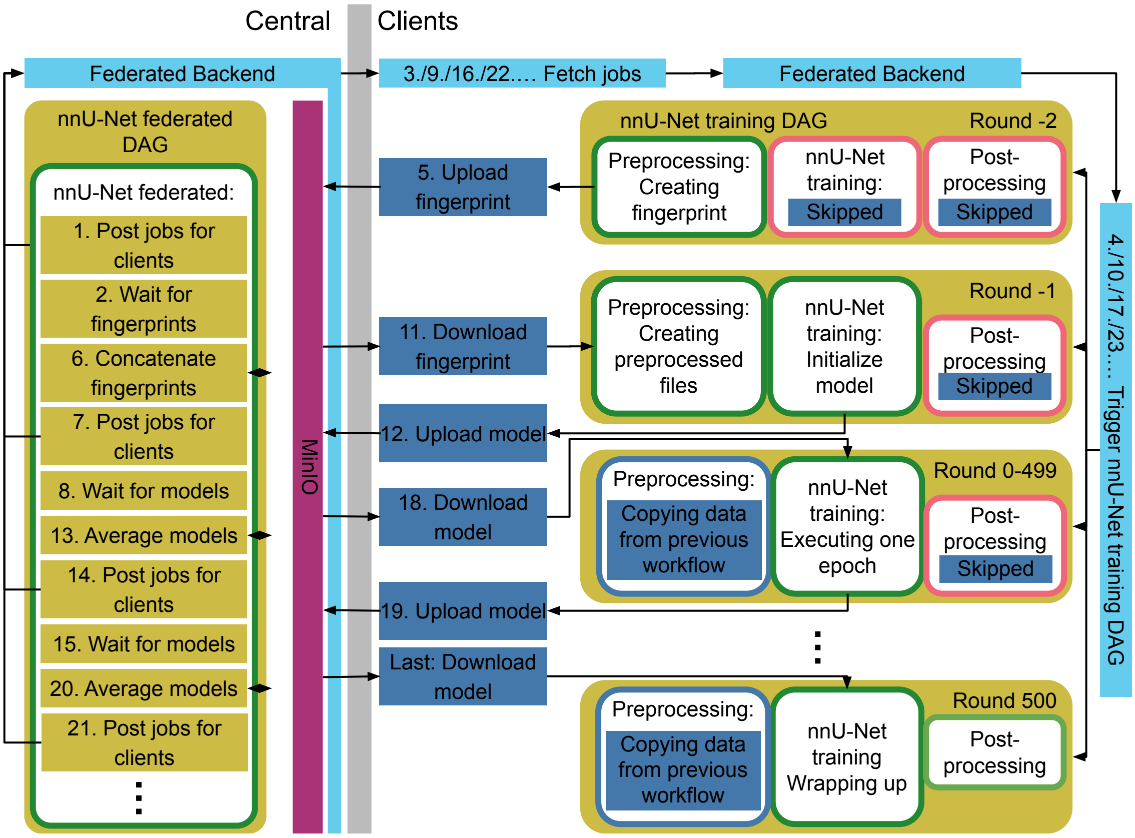


Figure 5.30: *Overview of the federated nnU-Net training using Kaapana. The central instance on the left side consists of a federated backend, MinIO and, the nnU-Net federated operator. The client instances on the right side consists also of the federated backend, which is responsible to trigger the client nnU-Net training DAGs. The nnU-Net training DAGs is represented by a simplified version of the actual training DAG, which consists of more than three operators. The yellow boxes correspond to Airflow DAGs with their operators. All dark blue boxes describe the pre- or post-hooks of the operators. The numbers indicate the order of processes during federated learning. This figure was adapted from Kades et al. (2022b) with permission of Springer. DAG=Directed Acyclic Graph, Kaapana=Kaapana is an open-source toolkit for the state-of-the-art platform provisioning in the field of medical data analysis, MinIO=High Performance Object Storage, nnU-Net=no new net U-Net.*

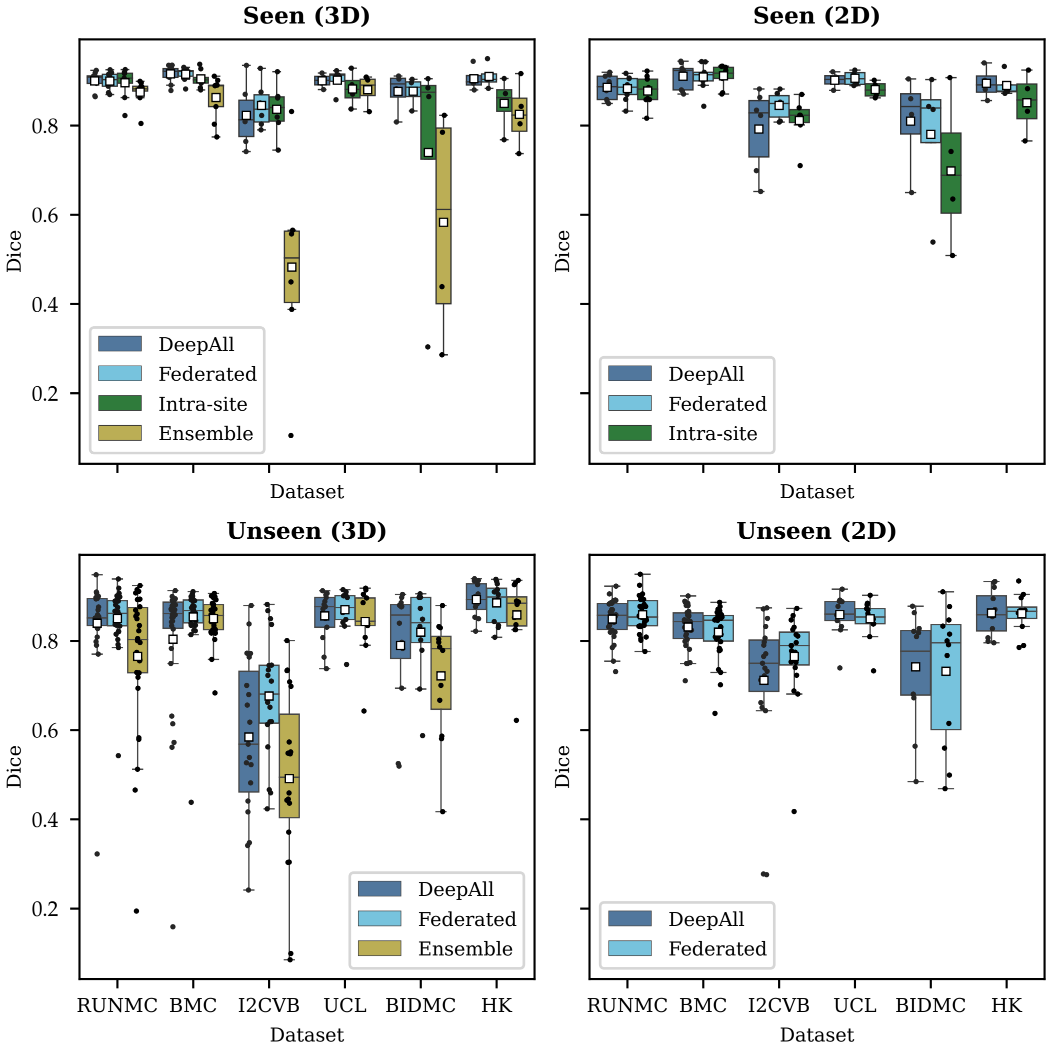


Figure 5.31: *Distribution of DICE scores for the different algorithms and datasets. For each site, the lower and upper quartile, and the median position is illustrated. The shape of the distribution is visualized by the black points and the whiskers. The white boxes indicate mean DICE scores. The results of the 2D and 3D nnU-Net architectures are illustrated on the left and right sides. The results of the "seen" and "unseen" experiments are at the top and bottom of the figure. For the 2D experiments, it is not possible to include the ensembled performance, since not all single-site trained models are able to generate a valid prediction on the test images. The figures show that the algorithms struggle with some outliers, especially, in the "unseen" experiment. Furthermore, the variance in the DICE scores varies is different from site to site, indicating the diversity of data per site. This figure was adapted from Kades et al. (2022b) with permission of Springer. DICE=Sørensen-Dice coefficient, nnU-Net=no new net U-Net.*

tape archive (TAR) archive object to the MinIO object storage using a pre-signed URL that it received with the job (5.). It is worth noting that all requests to MinIO need to pass the federated backend first since the MinIO endpoints are not white-listed for external communication. In the central instance, the collected fingerprints are merged into one fingerprint, describing all datasets of the participating sites (6.). In the next round, the fingerprint computed in the central instance is downloaded by a pre-hook of the preprocessing operator (11.). In the preprocessing operator, the data is preprocessed and the plan for the nnU-Net architecture is created. Using the same fingerprint at all client sites ensures that all clients configure the same preprocessing pipeline and model architecture. Once the image data is preprocessed, the weights and biases of the nnU-Net model are initialized with the nnU-Net training operator and uploaded to the central instance (12.). In the central instance, the model weights and biases are averaged using the $\mathtt{Fed}Avg$ method (13.) and the first training round for the client sites is triggered. For the following training rounds, the procedure is the same: First, the preprocessed data from the previous run are copied using a pre-hook of the preprocessing operator into the current working directory of the DAG run. Second, the averaged model is downloaded by a pre-hook from the central instance to the client instance and then used for training for the duration of one epoch in nnU-Net training operator. In a post-hook, the model is compressed again into a TAR archive object and uploaded to the central instance to be combined again with the other models. After training for 500 epochs, the averaged model is downloaded. In a postprocessing step, the trained model is stored on the local instances and training reports are generated.

Experiment results

Table 5.15 shows results for both, "seen" and "unseen", setups. It reports the DICE (%) and the ASD (mm) on the respective test datasets for the 2D and 3D–$\mathtt{full}res$ nnU-Net models, trained using Intra-site, Ensemble, DeepAll and Federated approaches. Additionally, the baselines from Liu et al. (2020a), Liu et al. (2021) and Gu et al. (2021a) are included. However, it should be noted, that the DCA-Net baseline uses different training, validation and test splits. The ensemble performance is only reported for the 3D models because the single-site 2D models could not generate a valid prediction on all target domain cases, belonging to an "unseen" site. The reported DICE scores are calculated by taking the arithmetic mean over the DICE scores of all test cases per site. The DICE scores reported in the "average" column are calculated using the unweighted average over the per-site mean scores. As in the previous section, a ranking analysis is also applied. In detail, a case-based ranking of the models is created per dataset. By taking first the mean ranking over the cases and then the mean over all different sites, an average rank per method is calculated and reported in the column "rank". The ranks are calculated respectively for the "seen" and "unseen" experiments but across the different architectures and algorithms. A significance test using the Mann-Whitney U test, Benjamini-Hochberg corrected, shows that in all experiments the differences in DICE scores between the centralized and federated trained nnU-Nets are not significant (all $p > .05$). In addition to the scores in Table 5.15, the DICE scores and the ASD of the cross-site performances along with STDs of the individually trained models are reported in the appendix in Table B.4 and Table B.5.

In the box plots of Figure 5.31, the distribution of the case-wise DICE scores for all datasets and methods are shown. The figure shows that in all datasets certain outlier cases exist, which are hard to segment. In Figure 5.32 the training loss of the centralized and the averaged training loss of the federated trained nnU-Nets are illustrated. For the 3D models, the averaged federated loss is slightly lower than the centralized loss. However, the effect is not reflected in the corresponding DICE scores.

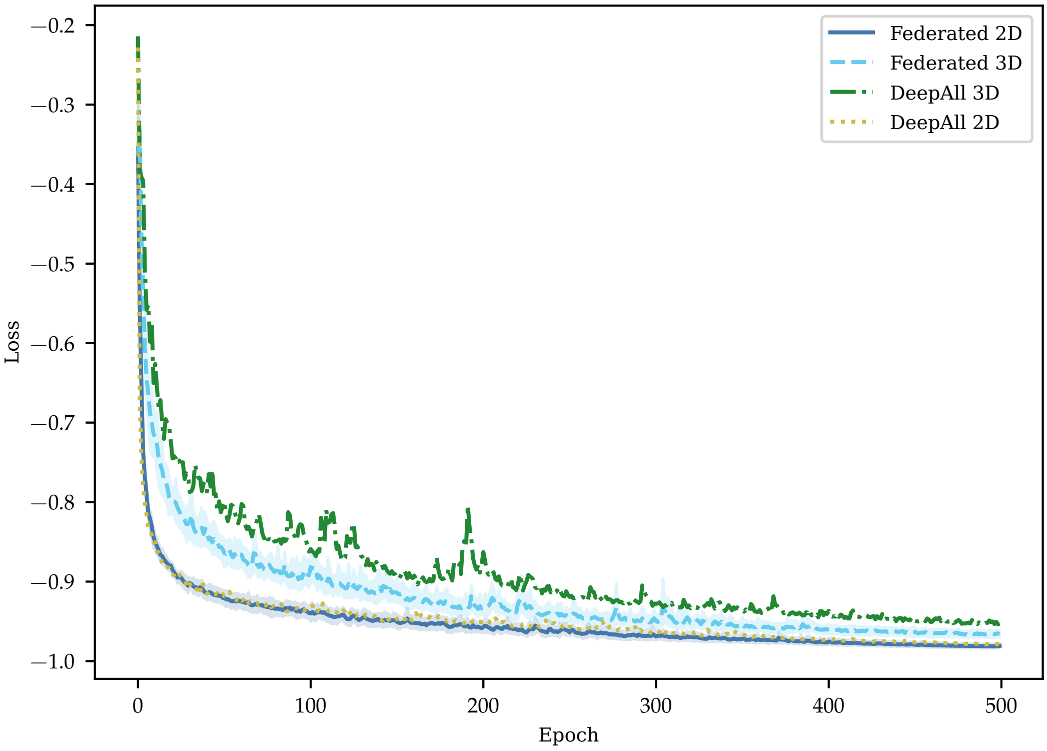


Figure 5.32: *Loss as a function of the training epochs for the "seen" experiment setup. Curves for the centralized and federated trained 2D and 3D nnU-Nets are shown. The loss curves and shaded areas of federated models are the mean and standard deviation over the losses at each of the six sites per epoch. While the loss for the 2D models is relatively similar between the centralized and federated training, the loss of the 3D federated model is by a small margin lower than the one of the 3D centralized model. However, the lower loss has no significant effect on the DICE score as seen in Table 5.15. This figure was adapted from Kades et al. (2022b) with permission of Springer. DICE=Sørensen-Dice coefficient, nnU-Net=no new net U-Net.*

6 Discussion

6.1 Semantic modeling for semantic textual similarity

6.1.1 Approach 1: Voting regression

In general, the results on the test set are slightly higher than on the train set for all approaches. This is also the case for the ClinicalBERT baseline. The Enhanced BERT shows no improvement compared to the baseline, indicating that the additional knowledge given to the CLS token token for the input to the classification layer has no visible effect.

The Voting Regression approach shows an improvement on the validation set. However, on the test set the performance decreases in comparison to the baseline, suggesting an overfitting on the training set. Another reason might be the imbalances between the training and test dataset (Kades et al., 2021)

6.1.2 Approach 2: $\mathcal{M}$-Heads

Introducing $\mathcal{M} = 4$ heads improves the performance on the validation and test dataset compared to the baseline performance. Especially, for the test set a higher PCC is measured. An explanation of good performance might be again the imbalances between the training and test dataset because each head might specialize in different characteristics of the training dataset, which can then be exploited during testing.

Set	Sentence a	Sentence b	T	A1	A3
Training	Ondansetron, 4 mg, 1 tablet, three times a day	Amoxicillin, 500 mg, 2 capsules, three times a day	3.00	1.68	1.70
Training	Prozac, 20 mg, 3 capsules, one time daily	Aleve, 220 mg, 1 tablet, two times a day	0.50	2.02	1.68
Training	Hydrochlorothiazide, 25 mg, one-half tablet, every morning	Ibuprofen, 600 mg, 1 tablet, four times a day	1.50	1.59	1.70
Test	Aleve, 220 mg, 1 tablet, two times a day	Acetaminophen, 500 mg, 2 tablets, three times a day	1.50	2.74	1.68
Test	Lisinopril, 10 mg, 2 tablets, one time daily	Naproxen, 500 mg, 1 tablet, two times a day	1.00	2.29	1.69

Table 6.1: *Comparison of the similarity scores as predicted by the Voting Regression (approach A1) and the medication graph (approach A3) along with the reference similarity score (T) for the corresponding sentence pair. The sentences are randomly selected examples and only the relevant entities of the original sentence are listed. This table was adapted from Kades et al. (2021).*

6.1.3 Approach 3: Medication graph

For the evaluation of the medication graph, scores prescribing medication were replaced in approaches 1 and 2 by updated medication graph scores. For both approaches, almost no improvement on the validation set is recognizable, however, on the test set, performance gains are noticeable. The best score on the test set of a PCC of 0.883 is created by combining all approaches. To further evaluate the effect of the medication graph, the difference in the MSE for Approach 1 and Approach 2 only on the subset of the medication sentences is calculated, yielding a MSE of 0.70 for approach 1, and with the medication graph a MSE of 0.58. The improvements by the medication graph approach as well as the cluster analysis of the baseline scores from Section 4.1.1 showcase that BERT struggles with domain-specific knowledge, and additional knowledge is necessary to cope well with these domain-specific sentences.

The only marginal improvements on the validation set in comparison to the performance increase on the test set can have multiple explanations. One explanation is again an imbalance between the training and test set. In detail, the distribution of labels in Figure 5.1 shows that the test contains more sentences with a lower label rank than the training set. Particularly, for the sentences prescribing medication lower scores in the test set are observed. i.e., for sentences prescribing medication, the mean and STD of the scores in the training set are 2.03 and 1.05, respectively, in comparison to the mean and STD of 1.10 and 0.50 on the test dataset. A complementary observation is the tendency of the medication graph to dampen the predictions, i.e. to lower the input scores. For the 94 medication sentences in the test set, the mean prediction score of 2.58 was lowered to 1.78 after applying the medication graph approach. Table 6.1 shows examples of sentences for which the medication graph altered the scores. The tendency to lower the scores for the medication graph might have two main reasons. Firstly, low scores on the edges (1.87 on average) in the medication graph are observed. Secondly, the incorporated formula of calculating the resistance of parallel circuits enforces a low final score in case there is at least one edge with a low score involved. The facts that the labeled scores in the test set are on average lower than in the training set in combination with the tendency of the medication graph to dampen scores, seem to be the major reason for the performance gain of the medication graph on the test set. Finally, the much lower percentage of sentences prescribing medication in the training set (147 out of 1642, 9%) than in the test set (94 out of 412, 23%) is another indicator of why only a neglectable effect is present in the performance of the validation set.

6.1.4 Limitations

One main limitation is that the proposed methods are only evaluated on the presented dataset, which has its own, unique characteristics. The success of the methods also varies between the validation and test dataset. The variance is mainly due to the imbalance between the test and training data set as illustrated in Figure 5.2, showing that the types of sentence pairs in the test dataset are only a subset of the types in the training set. Also, the differences in the label distribution and the number of words per dataset from Figure 5.1 are responsible for the inconsistent performances on the validation and test dataset. To benchmark the significance of the different approaches a more profound evaluation is necessary. Especially the usability and efficacy of the medication graph have to be further investigated either with more datasets containing medications or on datasets from other domains where extrapolation of information from known entities is necessary and where this information is not directly computable.

6.2 Semi-structured data analysis for text summarization

6.2.1 Automatic evaluation

The automatic evaluation of all methods using the ROUGE-1 F1 scores given in Table 5.4 shows that the proposed approaches to improve the BERT2BERT baseline lead to a significant improvement. Interestingly, the combination of both methods does only add a marginal improvement. Both approaches, the pointer and the extractive approach build on the idea of copying words or whole sentence structures from the general information and findings section into the conclusion. Taking a look at the example in Table 5.7, the effect becomes clear by the generation of the correct date. The example also shows that the pointer and the extractive approaches create more phrases constraint to the source input in contrast to the BERT2BERT model, which predicts more new phrases.

The success of the pointer and the extractive approaches might suggest, that also physicians might copy passages in their daily praxis between the two sections.

The analysis per TRC shows that models performed best on the reports with the CR class. Reasons for this might be the huge number of training samples (almost one-third of the reports) for this category, but also the uniformity of the template and that for healthy patients not much important information needs to be extracted from the findings. Furthermore, it has to be noted, that for the classes PR and PD, in some cases a tendency is mentioned (cf. Section 3.2.2), making it more difficult for the models to generate the report correctly.

The results of the key extraction method presented in Table 5.3 show that the Longest-k method generates the highest overlap between the extracted sentences and the conclusion, suggesting that the longest sentences contain the most important information for the conclusion. The importance of the extraction method of salient sentences is highlighted by applying the random baseline for the extractive approach. For the BERT2BERT extraction model an improvement in the ROUGE-1 F1 is noticeable. In contrast, in the hybrid model, combining the extractive and pointer approach, random extracted sentences only marginally influence the final performance.

The results of the automatic factual correctness check given in Table 5.5 suggest moderate factual correctness, except for the PR class. The main reason behind this is most probably the imbalances in the class distribution.

6.2.2 Expert evaluation

The human evaluation per TRC from Figure 5.5 show that for reports with the label CR, all generations show relatively high scores between 4 and 5. Also, for reports with label SD, the scores for the generated conclusion of the extractive approach are relatively close to the reference. For reports with the class PR and PD, the generation

seems to be more challenging. Probably, due to the more complex findings and the small number of training data for those classes. The comprehensibility of almost all generated conclusions is relatively high in comparison with the reference.
The rank distribution in Figure 5.6 shows that during the ranking of the generated conclusions, almost all models are capable of ranking first but also last for a few cases. Interestingly, for the comprehensibility criteria, the reference only ranks in the middle position. However, considering the broad distribution of ranks, the analysis emphasizes that based on the limited number of human-evaluated samples it is relatively hard to pin down a clear winner. The ranking stability plot in Figure 5.7 over the different evaluation criteria supports this observation. Although, the upper baseline with the reference and the lower baseline with the BERT2BERT model is recognizable.

6.2.3 Limitations

A major limitation is that the presented methods are only applied to SORs. It is an important next step to also evaluate the methods on FTORs or data from other domains. Due to a less accurate structure, new challenges might arise when applying the methods to FTORs. A further limitation is the design of the experiments. Therefore, in the future, it might be necessary to incorporate cross-validation and hyperparameter optimization in the development of the models with the target to obtain the best model configurations and to report a confidence interval. Moreover, due to time-consuming annotations, human evaluation is very limited in this study. Therefore, in total only 20 reference conclusions were compared with generated conclusions. While this is sufficient for a rough assessment of the models, it is limited to significantly evaluating the proposed models as the ranking distribution of the approaches in Figure 5.6 has shown. The human evaluation itself could also be improved, i.e. by forcing the annotators to rate one generation better than another one, by increasing the number of annotators and by examining the inter-annotator agreement. Also, it has to be noted, that the human evaluation has been only executed on the, according to the ROUGE score, best reports. A human evaluation of randomly selected reports might give better insights into the performance differences between the different presented approaches. Finally, the presented models do not take into account the imbalances in the TRC of the datasets. The performances of the models might be improved when incorporating this knowledge.

6.3 Assessing distributional shifts for text classification

6.3.1 The challenges of the dataset

It is a nontrivial problem in NLP to predict oncologic outcomes from FTORs using machine learning since the detection of disease progression relies on temporal and contextual reasoning rather than extracting specific information from a radiology report such as particular diseases or conditions (Fink et al., 2022a; Pons et al., 2016; Weber et al., 2020). Especially, when dealing with a distributional or domain shift as the Clinical TempEval 2017 challenge has shown. The challenge results showed a 0.20 F1 score drop in performance across domains, when training a NLP model on one cancer domain, e.g. colon cancer, and predicting timelines in another cancer domain, e.g. brain cancer with maximum F1 scores between 0.51 and 0.59. Cite
The lexical complexity as well as the characteristics of the patients of the three used datasets presented in Tables 5.9 and 5.10 as well as in Figure 5.10, shows that the experiments are executed on reports of an entire oncologic spectrum. The reports differ in the distribution of oncologic diseases, symptoms, and described procedures. While the reports in SORs and FTOR-DKFZ cover a relatively wide spectrum, the reports in FTORT-TKH describe only a specific spectrum of diseases. The reporting style varies strongly between SORs and FTORs, but also radiologists interpreting the different cancer types make a variety of linguistic choices and use their clinical jargon when discussing the oncologic findings (Fink et al., 2022a). In addition to the content-related differences in the reports, the lexical complexity and the lengths of the reports also vary strongly between the datasets. In comparison to public datasets from other domains, the medical datasets feature similar lexical characteristics. However, the used vocabulary of the reports is less known to the BERT model than the vocabulary of the public datasets, forcing the BERT model to split words multiple times into smaller known units.

6.3.2 The difficulty of a correct tumor response category assignment

The difficulty of assigning the correct TRC to a report is not only reflected in the only moderate F1 scores of the different baselines but already in the data selection and reference annotation process. The concept of the SORs suggests that dedicated TRCs should only be used in the absence of equivocal findings. In the case of equivocal findings, radiologists are encouraged to use narrative text to articulate ambiguities instead of adhering to the defined terminologies (Fink et al., 2022a; Weber et al., 2020; Eisenhauer et al., 2009). Therefore, using a regex to automatically extract the four RECIST-related TRCs to automatically generate a reference annotation for the SORs, already 3180 reports dropped out from the original 13685 SORs because no TRC could be extracted. While this dropout significantly decreased the number of training data it also ensured high-quality training data, which is a prerequisite for the development of good-performing NLP models (Willemink et al., 2020; Fink et al., 2022a). Similarly, from the original 884 FTORs, 82 reports had to be removed because no TRC could be assigned. While the final dataset contains a TRC label for all reports, the only moderate inter-rater reliabilities support the evidence of previous surveys, that many clinicians, even experts, struggle with the clarity of reported findings in radiology reports (Fink et al., 2022a). Therefore, in the reference annotation, the radiologists reached a Cohen's kappa of 0.77 and 0.90 for the FTOR-DKFZ and FTORT-TKH, respectively, indicating already the more difficult to read FTOR-DKFZ. Already in the reference annotation for the FTOR-DKFZ, the most difficult TRC is the SD (cf. Figure 5.11). The measured Cohen's kappas between the annotators of the human baseline from Figure 5.12 emphasize again the difficulty of a unique TRC, with relatively low Cohen's kappas between 0.21 and 0.58 on the FTOR-DKFZ and 0.31 and 0.79 for FTOR-DKFZ.

6.3.3 The human and machine baselines

The main idea behind the presented experiments is to use the labeled knowledge of the SORs and to automatically label with this knowledge FTORs. The machine baselines of the experiments in this work show a similar drop as in the Clinical TempEval 2017 challenge with the BERT model reaching F1 scores of 0.85 on the SORs but only 0.67 and 0.72 on the FTOR-DKFZ and FTORT-TKH, respectively. The distributional shift is handled surprisingly well by the Linear-SVC, reaching almost similar performance as the BERT model on the FTOR-DKFZ, whereas on the FTORT-TKH and SORs a drop in performance is observed. The good performance on the FTOR-DKFZ might be explained by the long reports, giving the TF-IDF and the Linear-SVC more information than in the case of the shorter FTORT-TKH and SORs. For the BERT model, the long reports of the ftorḍkfz pose a challenge, because in its self-attention it has relatively long sequences and it has to filter out the most relevant information without any direct corpus knowledge like it is present in the TF-IDF model.

The human baseline performances of the three user groups show varying performances. On respectively the FTOR-DKFZ and FTORT-TKH, the radiologists perform best with F1 scores of 0.74 and 0.84, followed by the students with 0.67 and 0.80 and the RTs with 0.56 and 0.74. Also for the human baseline, a drop in performance between the FTOR-DKFZ and FTORT-TKH is recognizable. Probably, due to the longer and therefore more complex in the FTOR-DKFZ dataset.

Evaluating the performances of the NLP models for the different TRCs using the AUC score, the Linear-SVC shows comparable results to the BERT model on all datasets. While it performs slightly below the BERT model on the FTORT-TKH it outperforms the BERT model by a small margin on the FTOR-DKFZ. Suggesting that at least for tasks with a huge distributional or domain shift between the training and testing data, a baseline of conventional, non-transformer-based NLP models is an important building block, when tackling the task of text classification. The corresponding ROC plots illustrate again the already mentioned performances between the three annotator groups. Similar to the observation in the text summarization tasks in Section 6.2, the analysis show, that both, humans and models, struggled most with the reports of class SD. One major reason for the difficulty on the SD is the ambiguity of the oncologic descriptions and the diverging meanings between the descriptions given in the findings and the interpreting radiologists final impression of disease progression (Fink et al., 2022a). In its definition, the RECIST category SD comprises a wide range of subthreshold changes in tumor burden ranging from formal disease progression to partial response (Fink et al., 2022a). e.g. in one of the misclassified FTOR, whereas the findings section refers to increasing lesions, the impressions section states "stable disease with a trend toward increasing tumor burden".

The ranking analysis in Figure 5.19, as well as the per annotator scores in the appendix in Table B.2, highlights that also within the three annotator groups, huge differences in performances are recognizable. Therefore, Student 1 reaches similar performances as the radiologists, whereas RT 3 performances are worse than those of the other RTs. The performance variations can be explained by the different career stages of the annotators and differences in German proficiency. The BERT model and Linear-SVC group themselves somewhere between the RT and the students.

The results of the hyperparameter optimization in Figure 5.9 show that the transformer-based models require much less optimization than the statistical-based methods, where parameters such as the threshold for a minimum document frequency have a strong influence on the overall performance. For the transformer-based model, the most relevant variable seems to be the learning rate, which should not be chosen too high.

The considerations of interpretability in Section 5.3.2 show that weights assigned by the TF-IDF or the attention weights of the BERT model might help to qualitatively mark tokens to indicate a certain decision of the algorithm. However, the quality of the assigned weights should be examined quantitatively in the future. Moreover, the UMAP representation of the CLS token embeddings shows that the learned representations after fine-tuning incorporate not only knowledge about the class, but also reflect the length of the report and the distributional shift between the different used datasets. Whether this behavior limits the performance of the classification and methods to create more universal representations might be the subject of future work.

6.3.4 The importance of confidences

The analysis of the confidences per TRC of the different annotators in Figure 5.13, shows that depending on the level of expertise, also the level of confidence increases. The confidences of the machines in Figure 5.15 show that some algorithms tend to give on average higher confidences. The BERT model assigns on average the highest probability to a class. The plots which illustrate the calibration of the confidences in Figure 5.14 indicate that for both FTORs, the annotators slightly underestimate their performance, especially, when assigning small confidences. The confidences of the machines-based models illustrated in Figure 5.11 create relatively well-calibrated curves. Except for the FTOR-DKFZ at smaller confidences the model overestimates their performances, however, this could also be explained by the low number of samples for the lower bins. Table 5.11 underlines that the probabilities created by the machine models are better calibrated than those of the human annotators.

How and why calibrated confidences are important is showcased in Figure 5.20. The figure presents the tumor-burdon change on a patient level over time. This kind of representation could be, e.g., helpful in tumor board assessments. The indicated confidences are of importance for the radiologists whether to trust the given data point or not.

6.3.5 Combining human- and machine-generated annotations

If ensembling techniques lead to performance improvements and if machine-based annotations could improve human-based annotations is illustrated in Figure 5.17.

The figure shows that ensembling the annotations across different annotations leads in all cases to higher scores than those reported when calculating the F1 score using micro-averaging. Furthermore, comparing the annotations of a single annotator or model (white background) to the combined ones, which include the RTs, students, radiologists, and machines (grey background), the performance is not always increased. This mostly depends on how much the performance varies between the different combined annotations. For example, for the FTORT-TKH, combing the Linear-SVC and BERT model predicted annotations leads to a loss in performance, mostly because of the poor performance of the Linear-SVC model.

Ensembling the human annotations with the ones from the machines (BERT and Linear-SVC), the performance of almost all annotators is improved (dark blue markers) for the FTOR-DKFZ. However, for the FTORT-TKH, often performance losses are observed. An explanation might be again the poor performance of the Linear-SVC model on the FTORT-TKH. Nevertheless, the performance improvements of the RTs for the FTOR-DKFZ using ensembling techniques show that the annotations of a human annotator with limited medical experience could be enhanced using machine-based annotations.

6.3.6 Correlations between dataset characteristics and performances

How different characteristics of the dataset influence the performances of the models and humans are presented in Section 5.3.2. Plotting the confidences of the annotators and the two machine-based models in Figure 5.21 shows that the confidences of humans are reflected in the performance of all models and human annotators. While this correlation is less profound for the confidence of the BERT model it almost vanishes for the Linear-SVC. The analysis shows that transformer-based models might be better suited when trying to create human-like confidences. The correlation plots between the lexical complexity and the performances showcase, how specific lexical characteristics of the dataset influence the results of the human- and machine-based approaches. This knowledge could be used to optimize the performance of radiologists by deliberately showing them more difficult reports, such as long reports during training. Similarly, when training the machine further losses or over-sampling of certain reports during training could be introduced. From a medical point of view, discrepancies between oncologic and non-oncologic findings lead to substantially worse performances of the human annotators and the NLP models as illustrated in Figure 5.25. Those discrepancies are created by divergent semantic tendencies such as progressive disease and improvement, stable disease and worsening or partial response and worsening. As discussed before, interestingly, while the confidences of the machines are significantly influenced by the discrepancies in oncologic and non-oncologic findings, the confidences of the human annotators do not significantly change (P_{value}=.38).

6.3.7 Limitations

The presented experiments have multiple limitations. First, since the TRC is automatically retrieved from SORs and not manually reviewed, the quality of the assigned TRC and whether the labels follow the RECIST-related SOR concept is unknown. However, due to a four-eyes principle and final approval by an attending radiologist, the extracted labels should be considered as an appropriate reference annotation (Fink et al., 2022a). Second, due to a high expected workload, the reference annotation for the FTORs is only done based on the report itself, not including any images from the radiologic examinations. Furthermore, in contrast to the SORs, the FTORs do not contain tables with reference measurements of the target lesions. Consequently, the created predictions of the annotators and NLP model do not represent quantitative RECIST measurements and classifications (Fink et al., 2022a). Third, it has to be pointed out, that the human annotators all possess a certain medical understanding due to their educational background in contrast to the pre-trained BERT model or the TF-IDF-based models. This disadvantage of the machine-based models might automatically lead to a slight performance decrease in comparison to human annotators. Contrary, the machine-based models have the advantage of seeing all SORs during training. One subject of future work could be to prime the BERT model more on the target dataset using transfer learning approaches such as further pretraining tasks on medical datasets or unsupervised training tasks on the target dataset. Fourth, in the design of the experiment, also during the predictions, only the general information and findings sections are used as input for the classifier, including also the reference could drastically improve the performance of the human annotators and the NLP models. Fifth, all experiments are only executed on German SORs, limiting the generalizability of the results. Sixth, to handle the imbalance of the TRCs in the data, a weighted cross-entropy loss during training was introduced. The effect of this adaptation as well as further approaches to handle imbalanced data such as introducing a focal loss or a class-weighted sampling should be examined in future work in more detail. Seventh, the memory consumption of transformer-based models scales with $\mathcal{O}(n^2)$ with n the sequence length, enforcing techniques to handle the often long medical reports. The splitting and mean pooling of the report, applied in this work represent only a transition solution. Further existing solutions should be implemented and investigated. Eighth, the qualitative analysis of attention weights of the CLS token token is very limited and should be done in more depth in the future. However, the visualization suggests, that a potential improvement of the text-level interpretability could be achieved by removing the noise of the pre-trained attention weights from the attention weights learned during fine-tuning.

6.4 Manual data annotation and tagging for medical images

6.4.1 The implementation

The approach connects two existing open-source projects by embedding an iframe of the OHIF Medical Image Viewer into Doccano. The implementation allows besides fast tagging and editing of DICOM tags, also the possibility to add free-text to DICOM images. The feasibility of different user scenarios including the definition, curation and annotation of cohorts for their application in subsequent medical image analysis workflows is demonstrated by a detailed description of the necessary steps. While the presented implementation mainly represents a proof-of-concept it could be extended in multiple ways to a production-ready system (Kades et al., 2022a). Replacing or complementing the current iframe solution with more native javascript-based tools like dcmjs (dcmjs, 2023) could extend the variety of tasks concerning text-based annotations even on a pixel-level (Kades et al., 2022a). For example, the sequence labeling task of Doccano could be used to link a bounding box to a text sequence as it is needed for different image-text correlation algorithms. The range of applications goes beyond the presented use case scenarios. e.g., after the application of an algorithm on DICOM images, it could be used to assess the quality of the results or additional metadata could be added to DICOM images.

6.4.2 Limitations

A major drawback of the current implementation is, that the OHIF Medical Image Viewer visualizes images based on the Study Instance UID Attribute, which prevents the labeling or annotation on patient, series or even slice-level, which might be needed in certain use case scenarios. A further limitation of the presented implementation is the assessment of its usability and the speed at which annotations can be done. While the Doccano tool itself is optimized for a fast tacking of texts, it would be interesting to benchmark the time needed for the annotation of DICOM images.

6.5 Real-world federated learning for the task of image segmentation

6.5.1 The implementation

To enable federated learning using Kaapana multiple adjustments and extensions are presented in this work. The broader goal behind the implementation is to provide building blocks that not only allow federated learning scenarios, but also all kinds of communication and file transfer between instances. Therefore, among the two most important features is the ability to trigger a workflow on a remote instance and transfer files created by Airflow operators to the object store of another instance. The design choice allows the implementation of all types of network topologies, such as decentralized or hierarchical, and federated computing plans, such as peer-to-peer or sequential (Kades et al., 2022b). Furthermore, a single execution of a remote running workflow and a subsequent collection of the created results is possible. Given an existing implementation of a segmentation algorithm, a lot of libraries for federated learning such as OpenMined (OpenMined, 2023) require many adjustments to the code for a federated training. In the implementation presented here, only very few adjustments to the training pipeline are necessary, since in between federated training rounds, the training pipeline is interrupted and the weights are updated on a file level. Therefore, the only requirements are the interruption of a training pipeline and a file-based storage of the latest checkpoint. While the complete interruption of the training can be time-consuming, any existing pipeline can be adapted with moderate efforts for federated learning scenarios. Using the implementation, Figure 5.30 shows how federated training of the nnU-Net with federated averaging can be set up without changing anything in the locally running nnU-Net workflow.

6.5.2 Experimental results

The results from the various experiments demonstrate the potential of a federated training of the nnU-Net using federated averaging. The federated trained model achieves similar or even slightly better performances to a centralized trained model, eliminating the need to pool data for training. The slightly, but not significantly, better results could be explained by the fact that the federated approach might have a regularizing effect on the training. In addition, the federated approach does not take into account the slight imbalance in the number of cases between sites, whereas the centralized approach might be biased towards sites with a higher number of training data. Considering a centralized approach as an upper baseline, the good performance of the federated averaging approach raises the question of whether more sophisticated averaging or aggregation strategies could lead to notable potential improvements for a federated training of the nnU-Net. Comparing the performance of the model ensemble to the centralized and federated trained models, performance gaps are visible for some datasets. Although auxiliary results with the nnU-Net do not show signs of overfitting, the performance gaps could be explained by the relatively small number of training samples at some sites, resulting in models that highly fit the data of the sites and generalize worse to other sites. This could make the ensemble less stable.

In comparison to the provided state-of-the-art results, none of the presented models can compete, even not the centralized model. However, the nnU-Net is used without any further measures or adjustments in contrast to the state-of-the-art models, which are specifically designed to perform well on data from multiple sites. Furthermore, the scores reported for the DCA-Net cannot be compared due to inconsistent training, validation and test splits (Kades et al., 2022b). Figure 5.31 shows that for some outliers in the test dataset, the models struggled to generate good segmentations. The existence of outliers underpins the need for a consistent case-wise training and test dataset for a fair comparison of models, which is why the list of cases used for testing is given in the appendix in Table B.3.

The "seen" experiments show primarily that the intra-site trained models perform all moderately on their in-house data. At the same time, it is observed that training on more data indeed helps to surpass the performance of intra-site trained models in most of the sites. In general, it has to be noted that the results should be interpreted with caution because of the low number of test samples. Nevertheless, the good performance of the intra-site trained models might question the benefits of federated learning or the training on more heterogeneous data to obtain more robust models in the first place. Especially because of the technical and organizational difficulties coming with federated training. Of course, this question only arises if there is enough training data available at the site and also depends on the difficulty of the task.

The "unseen" experiments underline the results from Gonzalez et al. (2021); Full et al. (2021), that the nnU-Net model still needs improvements when applied to data from unseen sites to reach state-of-the-art performance. Therefore, measures need to be incorporated to make the model more generalizable and robust.

6.5.3 Limitations

Especially, in the medical domain, an implementation to enable federated learning scenarios needs to be optimized to work with huge data and models that require relatively long training times. In the case of the nnU-Net training times of up to two days and model sizes of multiple hundreds of megabytes are required. This prerequisite has the consequence of a bottleneck in the efficacy of one federated round. Although the design choice to start a new DAG for each federated round, makes the approach very agnostic towards the underlying model, the implementation is also very time-consuming, making it unattractive for applications that require many federated update rounds within a short amount of time. In those cases, federated communication could be incorporated directly within a processing container, however, still using the same setup for the communication of the results. Another limitation compared to existing frameworks is that many FL-specific algorithms are not implemented and that the privacy and data protection aspects are only partially covered.

A limitation of the experimental results is that it is only applied to one dataset. For a significant evaluation of the federated training of the nnU-Net, the performance has to be examined on more, real-world datasets, including datasets with different segmentation tasks and modalities. A further bottleneck of the nnU-Net itself for use in federated learning setups is its model size resulting in huge communication costs in each federated round. Therefore, interesting research topics include how to reduce the size of the model or how to reduce the number of federated rounds, i.e., by increasing the number of local epochs per federated round or investigating different federated aggregation or optimization methods.

7 Conclusion

This work discussed five specific challenges for the application of AI algorithm in multi-institutional clinical settings. While the problems only cover parts of the challenges and future steps presented in Section 1.3, the proposed solutions show great potential for future research directions and applications.

Semantic modeling for semantic textual similarity on medical textual data is an important subtask for multiple clinical applications, such as medical question-answering applications for clinical decision support, recognition of redundancies in medical reports, or, e.g., for the creation of a patient cohort with a similar diagnosis. Three approaches are presented in this work to tackle the problem of STS on medical data. In the first approach, BERT, is enhanced with traditional feature-based similarity measurements, the generated scores are combined with the similarities and processed by weighted regression models. In a second approach, M-Heads are applied to concentrate on different characteristics of the dataset. The third approach attempts to incorporate corpus knowledge by automatically extrapolating medical knowledge from the training data. All methods resulted in modest performance improvements compared to the BERT baseline, but the gains varied between the training and test datasets.

Possible future works include the evaluation of the approaches on new datasets from further clinical centers, from other domains, or in different languages. Especially, the potential and possible improvements of the medication graph need to be further evaluated on problems such as similarities between ontologies. Also, the incorporation of corpus knowledge into batch-wise training might add additional value to the training.

Finally, the M-Heads approach could be further investigated by, e.g., training a binary classification head, separately, for each class (Kades et al., 2021).

The semi-structured data analysis for text summarization proves that text summarization for medical reports is a promising research direction. The BERT2BERT-based abstractive text summarization model serves as a baseline for creating the conclusion of a German SOR based on the radiology findings. To improve the factual correctness and therefore the quality of the generated conclusions, two strategies are proposed. In the BERT2BERT+Ext the additional training task to reconstruct key sentences from the source input is introduced. Like this, the resulting summaries are closer and more concise to the findings sections in the medical reports. In the BERT2BERT+Ptr approach, the pointer mechanism is applied, which modifies the decoder's prediction to directly copy salient segments from the source sequence into the generated sequence. Despite imbalances in the TRC distribution in the training data, the two hybrid models greatly improved factual correctness by preventing the generation of unfaithful facts in the generated summaries. As already discussed in the limitations section 6.2.3, one major requirement to validate the efficacy of the presented methods is their application to free-text reports or datasets from other domains. Furthermore, the methods should be evaluated on texts in other languages than German. A thorough hyperparameter optimization could help to assess the methods from a more statistical point of view, as well. The factual correctness of the generations could be further improved by incorporating knowledge of the TRC classification algorithms presented in this work into the generation models, as it has been shown that the TRC classification of the hybrid models is still below that of the ones from the dedicated text classification algorithms. Moreover, the investigation of further pretraining or downstream training similar to the generation of key sentences might improve the summarization performance of the models. Since the reports are written based on an image and sometimes tabular data, the inclusion of those data might be another interesting research direction to create more concise summaries or even findings sections of the radiology reports. Finally, a more thorough human evaluation would help further assess the performance of the text summarization models.

Assessing distributional shifts for text classification is crucial when working with multi-institutional datasets. The presented work demonstrates the potential of using SORs as a data resource to facilitate automatic annotation of clinical FTORs, thereby reducing the time-consuming manual annotation effort of human experts. The presented system could be used to extract clinically relevant oncologic endpoints from large volumes of longitudinal FTORs which are crucial for automated clinical decision support for patients referred for multidisciplinary tumor board assessments (Fink

et al., 2022a). The state-of-the-art text-classification algorithms showcase that NLP models, trained on mined SORs, reach human-level performance in determining TRCs on FTORs and that non-expert and machine-based annotations can be improved by ensembling human and machine-based predictions. Moreover, the analysis shows that human and machine performance as well as certainties correlate to patient characteristics, lexical complexity, and semantic diversity of the radiologic reports.
In the future, the presented methods should be evaluated for the automatic information extraction of further clinically relevant properties such as diagnoses, lab results, measurements, or medical billing codes. Further analysis must be conducted on more datasets, including non-medical and non-German datasets, in order to assess the significance of performance differences on corpus-determined characteristics like lexical complexity. Also, the influence of further medical characteristics on the performance of the downstream task such as therapeutic exposures, tumor profiles, or the writer of the reports, represents interesting future work. Due to the high diversity in medical reports, a similar analysis would also be interesting for other downstream tasks such as STS or text summarization. Improving the methods to handle distributional shifts between training and testing data should in general be emphasized in future works. The information and knowledge gained from the subgroup analysis might help to customize and thus improve the performance and generalizability of the classification models. An example from this work is to focus on reports with oncologic and non-oncologic discordance in the training pipeline. Similar to the approaches for improving the performance of the text summarization models, the involvement of the images associated with the considered report might be beneficial for the classification task. Furthermore, incorporating corpus knowledge such as that learned by the TF-IDF analysis in a transformer-based model or a hybrid model consisting of a transformer and traditional NLP features could help to improve the classification performance. Moreover, applying unsupervised training tasks on the target free-text data might help to improve the domain adaptation ability of the presented NLP algorithms. Additionally, the interpretability and explainability of such NLP tasks deserve more focus in future works, because, especially in the clinical domain, the original reasons for specific predictions are of high importance. For example, in case of inconclusiveness of the algorithm text snippets for and against a certain label could be highlighted in a report. However, token-level explainability and interpretability of transformer-based models is still an object of research and should go beyond the qualitative analysis as presented in this work. Likewise, the application of transformer-based models on long sequences deserves more attention in the future.

Manual data annotation and tagging for medical images are crucial for multiple tasks when working on real-world multi-institutional datasets. Examples include the correction of false metadata, the creation of curated cohorts used for the application of AI algorithms such as the nnU-Net, or a retrospective assessment of generated results such as the quality of segmentations. The presented technical solution allows simple and fast manual data annotation and tagging of medical images by combining the OHIF viewer with the annotation toolkit Doccano (Nakayama et al., 2018), hosted using the Kaapana (Scherer et al., 2023) toolkit.

Motivated by the points presented in the limitation Sections 6.4.2, the efficacy in terms of time and usability of the manual annotation has to be assessed on sample real-world datasets in the future. e.g. the evaluation of computer-generated results could be compared to the implementation of Stein et al. (2019). Further future steps include support for pixel-wise annotations, bounding boxes, and measurements. For a high-level annotation of images, a kind of gallery view could be implemented that allows for even faster tagging and cohort creation. Also, the support of other image formats such as Nrrd or NIfTI or their conversion to the DICOM standard might include further improvements to the presented implementation. Finally, an incremental learning pipeline could also be easily tested and implemented since Kaapana comes with the possibility to train all kinds of AI algorithms.

Real-world federated learning for the task of image segmentation is a crucial step toward more robust and generalizable models that also perform on unseen data. The presented implementation allows the application of all kinds of federated learning strategies in real-world clinical settings. Kaapana as the base system provides the necessary infrastructure, data access, and preprocessing as well as the on-site execution of DL pipelines, which are often neglected in common federated learning frameworks. The federated training of the nnU-Net on the multi-site prostate MRI dataset and the comparison to single-site or centralized training provide important insights for the efficacy and necessity of federated learning. One insight is that only locally trained models may be sufficient to create well-performing data-tailored segmentation models in the presence of sufficient local training data. However, it is also shown that the federated trained model achieves equal performance to the centralized model and even boosts the overall performance of the models by a small margin.

Important future steps include the application of the proposed system in a real-world setting such as the RACOON project, which already uses Kaapana as infrastructure and aims to train a generic and robust segmentation model to automate and standardize the assessment of COVID-19-related tissue alterations (Kades et al., 2022b). The presented implementation can be improved in many aspects. e.g. the transfer and connection between sites could be extended by certificate-based authentication. Furthermore, common FL capabilities such as homomorphic encryption, encrypted computation, or differential privacy could be implemented. In parallel, it could be interesting to experiment with the integration and combination of existing federated learning frameworks, which focus more on the methodical aspects of federated learning. Besides the federated training of the nnU-Net, the inclusion of further federated training workflows and the application of algorithms to non-imaging data are promising future directions. However, implementations of various federated setups, such as peer-to-peer federated learning, are also of great interest. A significant future step for the federated training nnU-Net is the reduction of communication costs while maintaining centralized performance. Possible steps include reducing the size of the model and the number of federated learning rounds, as well as investigating various aggregation or federated learning strategies. Another major future research direction motivated by the results of the experiments is the need to improve the generalizability of the nnU-Net, especially, on unseen data that exhibit a distributional shift in comparison to the training data. Interesting approaches to the nnU-Net could be adapted from the works of Liu et al. (2021); Jiang et al. (2022); Full et al. (2021); Gu et al. (2021a). The high variety and heterogeneity of medical imaging data might also motivate the creation of multiple data-centralized models instead of only one generally trained model.

In summary, the approaches discussed represent promising steps toward the application of algorithms in multi-institutional settings. Key future steps include the rollout of the implementations and approaches into clinics to validate their practicability in the real world and to make use of the implementations for medical research questions. However, expanding approaches to include additional clinical data such as lab data or working with multi-modal data such as textual and imaging data in the method development process are promising future directions as well. Using the Kaapana project along with its associated projects such as RACOON, the presented approaches can be quickly deployed and evaluated on real and more heterogeneous clinical data, as well as used for medical research. The research environment allows for very short iterations to constantly work on improving the presented approaches on a technical and methodological level. When applying methods in a real-world clinical context, the research focus in the fields of interpretability and explainability will gain greater importance. Close cooperation between physicians, clinics, and developers will open up the evolvement of more standards on data handling and algorithms for a secure and generally valid evaluation and execution of algorithms in real-world multi-institutional settings.

A Own contributions and publications

This chapter gives a summary of my contributions and sets them apart from those of the full team. This thesis was written in the division of Medical Image Computing under the supervision of Prof. Dr. Klaus Maier-Hein, who is also the first supervisor of this thesis. I worked as part of a multidisciplinary team of scientists and collaborated with physicians and scientists from other departments and consortiums throughout the time.

The thesis is built on top of five publications, which are listed below in Section A.2 and to which various authors contributed. My individual contributions are distinguished from those of the other authors in the following section.

A.1 Own contributions

Contributions in first author publications

The publication, **Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study**, resulted from our participation in track 1, N2C2/OHNLP Track on Clinical Semantic Textual Similarity of the 2019 National NLP Clinical Challenges (N2C2). For the challenge participation and the writing of the publications, I was heavily supported by the co-authors of the paper. While I created the baseline and was mainly working on the first approach, "Enhancing BERT with features based on similarity measures", Jan Sellner was the driving force behind the "Medication graph" approach and Gregor Köhler behind the "M-Heads" approach. The data analysis and writing of the publication were shared between Jan Sellner and me.

The publication, **Fine-tuning BERT Models for Summarizing German Radiology Findings**, was created in the course of the master thesis of Siting Liang. The research question was proposed and driven by myself. Siting was responsible for most of the implementations and for the presented hybrid approaches. The used SORs were provided by Prof. Tim Weber. While I was the main supervisor and advised Siting on design and implementation decisions, Prof. Michael Strube and Prof. Klaus Maier-Hein helped additionally with conceptional questions. I was responsible for the whole annotation process. Annotations have been contributed by the following people: Matthias Fink, and Peter Full. In addition to Siting's analysis, I created the rank-based analysis of the human annotation for my thesis. The publication was written in the first place by Siting Liang and reviewed by myself.

The idea for the publication, **Deep Learningbased Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports**, was originally conceived by Prof. Jens Kleesiek. The FTOR-DKFZ was provided by Prof. Jens Kleesiek, the FTORT-TKH by Dr. Matthias Fink and the SORs by Prof. Tim Weber. All medically related content and interpretation of results were created primarily by Dr. Matthias Fink and Prof. Jens Kleesiek. The annotation process was coordinated by Dr. Matthias Fink and myself. Annotations have been contributed by the following people: Jens Kleesiek, Matthias Fink, Arved Bishoff, Martin Moll, Merle Schnell, Maike Küchler, Melina Fritz, Julia Jaworek, and Genoveva Wolf. All technical concepts, all implementations, and all statistical evaluations were developed and created by myself. The paper was primarily written by Dr. Matthias Fink, while I was responsible for all result figures, tables, and proofreading. The contents and results presented in this work go beyond those of the original publication.

The idea for the publication, **Efficient DICOM Image Tagging and Cohort Curation Within Kaapana**, was driven by the demand in RACOON to curate medical images.

The concepts, implementation, and use cases were developed by myself. The paper was primarily written by myself, with support from Jan Scholtyssek.
The concept behind the publication, **Towards Real-World Federated Learning in Medical Image Analysis Using Kaapana**, was also driven by RACOON with the target to train generic and robust models for a standardized and automated biomarker extraction. While the nnU-Net pipeline was already integrated for local use by Jonas Scherer, I was responsible for adjusting the nnU-Net for federated use cases and for all the concepts and implementation necessary to execute workflows in federated scenarios on Kaapana. The paper was primarily written by myself, with support from Jonas Scherer and Max Zenk.

Contributions to the Kaapana framework

It is worth mentioning that throughout my time working on the thesis, I was also heavily involved in the development, implementation, and design of the Kaapana framework presented in section 3.1.1. Besides general maintenance and core development tasks, I developed the user interface, designed and implemented the concept of extensions, which is similar to an app store and integrated many external services into the toolkit. I was involved in the development of processing pipelines and the integration of on-demand services and interactive processing pipelines. Furthermore, I provided many solutions to ease up the development within the platform. Through the development of Kaapana, I was able to use the framework for my benefit and to deploy some of my research directly into the clinics.

A.2 Own publications

This section lists all publications that I was a part of and that contributed to my work. It is subdivided into *First Authorships*, *Co-Authorships* and *Software*.

First Authorships - Peer Reviewed Journal Publications and Conferences

Klaus Kades, Jan Sellner, Gregor Koehler, Peter M Full, T Y Emmy Lai, Jens Kleesiek, and Klaus H Maier-Hein. Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study. In JMIR Med Inform, 9(2):e22795, Feb 2021. ISSN 2291-9694. doi: 10.2196/22795. URL https://medinform.jmir.org/2021/2/e22795.

Siting Liang, **Klaus Kades**, Matthias Fink, Peter Full, Tim Weber, Jens Kleesiek, Michael Strube, and Klaus Maier-Hein. Fine-tuning BERT Models for Summarizing German Radiology Findings. In Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 3040, Seattle, WA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.4. URL https://aclanthology.org/2022.clinicalnlp-1.4.

Matthias A. Fink, **Klaus Kades**, Arved Bischoff, Martin Moll, Merle Schnell, Maike Küchler, Gregor Köhler, Jan Sellner, Claus Peter Heussel, Hans-Ulrich Kauczor, Heinz-Peter Schlemmer, Klaus Maier-Hein, Tim F. Weber, and Jens Kleesiek. Deep Learningbased Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports. Radiology: Artificial Intelligence, 4 (5):e220055, 2022a. doi: 10.1148/ryai.220055. URL https://doi.org/10.1148/ryai.220055.

Klaus Kades, Jonas Scherer, Jan Scholtyssek, Tobias Penzkofer, Marco Nolden, and Klaus Maier-Hein. Efficient DICOM Image Tagging and Cohort Curation Within Kaapana. In Bildverarbeitung für die Medizin 2022, pages 279284, Wiesbaden, 2022a. Springer Fachmedien Wiesbaden. ISBN 978-3-658-36932-3. doi: 10.1007/978-3-658-36932-3_59. URL https://doi.org/10.1007/978-3-658-36932-3_59.

Klaus Kades, Jonas Scherer, Maximilian Zenk, Marius Kempf, and Klaus Maier-Hein. Towards Real-World Federated Learning in Medical Image Analysis Using Kaapana. In Distributed, Collaborative, and Federated Learning, and Affordable

AI and Healthcare for Resource Diverse Global Health, pages 130140, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-18523-6. doi: 10.1007/978-3-031-18523-6_13. URL https://doi.org/10.1007/978-3-031-18523-6_13.

Co-Authorships

Jonas Scherer, Marco Nolden, Jens Kleesiek, Jasmin Metzger, **Klaus Kades**, Verena Schneider, Michael Bach, Oliver Sedlaczek, Andreas M. Bucher, Thomas J. Vogl, ...Klaus Maier-Hein. Joint Imaging Platform for Federated Clinical Data Analytics. JCO Clinical Cancer Informatics, 4:10271038, November 2020. doi: 10.1200/CCI.20.00045. URL https://ascopubs.org/doi/full/10.1200/CCI.20.00045

Jonas Scherer, Marco Nolden, Jens Kleesiek, Jasmin Metzger, **Klaus Kades**, Verena Schneider, Hanno Gao, Peter Neher, Ralf Floca, Heinz-Peter Schlemmer, and Klaus Maier-Hein. Abstract: Joint Imaging Platform for Federated Clinical Data Analytics. In Christoph Palm, Thomas M. Deserno, Heinz Handels, Andreas Maier, Klaus Maier-Hein, and Thomas Tolxdorff, editors, Bildverarbeitung für die Medizin 2021, pages 127127, Wiesbaden, 2021. Springer Fachmedien Wiesbaden. ISBN 978-3-658-33198-6. doi: 10.1007/978-3-658-33198-6_31. URL https://doi.org/10.1007/978-3-658-33198-6_31.

Viktoria Palm, Tobias Norajitra, Oyunbileg von Stackelberg, Claus P. Heussel, Stephan Skornitzke, Oliver Weinheimer, Taisiya Kopytova, Andre Klein, Silvia D. Almeida, Michael Baumgartner, Dimitrios Bounias, Jonas Scherer, **Klaus Kades**, Hanno Gao, Paul Jäger, Marco Nolden, Elizabeth Tong, Kira Eckl, Johanna Nattenmüller, Tobias Nonnenmacher, Omar Naas, Julia Reuter, Arved Bischoff, Jonas Kroschke, Fabian Rengier, Kai Schlamp, Manuel Debic, Hans-Ulrich Kauczor, Klaus Maier-Hein, and Mark O. Wielpütz. AI-Supported Comprehensive Detection and Quantification of Biomarkers of Subclinical Widespread Diseases at Chest CT for Preventive Medicine. Healthcare, 10(11), 2022. ISSN 2227-9032. doi: 10.3390/healthcare10112166. URL https://www.mdpi.com/2227-9032/10/11/2166.

Maximilian Fischer, Philipp Schader, Rickmer Braren, Michael Götz, Alexander Muckenhuber, Wilko Weichert, Peter Schüffler, Jens Kleesiek, Jonas Scherer, **Klaus Kades**, Klaus Maier-Hein, and Marco Nolden. DICOM Whole Slide Imaging for Computational Pathology Research in Kaapana and the Joint Imaging Platform. In Klaus Maier-Hein, Thomas M. Deserno, Heinz Handels, Andreas Maier, Christoph Palm, and Thomas Tolxdorff, editors, Bildverarbeitung für die Medizin 2022, pages 273278, Wiesbaden, 2022. Springer Fachmedien Wiesbaden. ISBN 978-3-658-36932-3. doi: 10.1007/978-3-658-36932-3_58. URL https://doi.org/10.1007/978-3-658-36932-3_58

Software

Kaapana 0.1.3

Scherer J., **Kades, K.**, Gao, H., Floca R., Neher P., Nolden M., Maier-Hein K.
DOI: `https://doi.org/10.5281/zenodo.5786866`
URL: `https://github.com/kaapana/kaapana`

B Appendix

B.1 Annotation guides and screenshots

Different annotations were executed throughout this work. For all annotations, either an annotation guide was provided, or the annotation procedure was discussed before the actual annotation.

B.1.1 Annotation guide for TRC assignment of the FTORs

In the reference and human baseline annotation of the TRC for the FTOR-DKFZ and FTORT-TKH, in total, eight annotators were involved. Two radiologists (M.A.F. [in training] and J.K [board certified]) executed the reference annotations based on a common agreement. The seven annotators (A.B., M.M, M.S, M.K. M.F, J.J, and G.W) for the human baseline were provided with a short introduction by one of the reference annotators as well as with the following annotation instructions:

Annotation instructions for TRC assignmnet

Read through the report and give exactly one of the four classes to the finding:

- Regular (healthy patient): regular (r),
- Improvement of findings: improvement (b),
- Constant findings: constant (k),

- Deterioration of findings: deterioration (s),

and how certain you are about your statement:

- very sure (1),
- sure (2),
- neutral (3),
- uncertain (4),
- very uncertain (5).

The letters/numbers in brackets are key assignments with which you can select the class. You can change the report with the left/right arrows. Make sure that at the end of each report, exactly one diagnostic class and one scale of certainty has been given.

B.1.2 Annotation guide for expert evaluation on generated conclusions

The three main criteria to judge the generated systems were determined in agreement with domain experts. The two annotators were presented with the following annotation instructions.

Annotation instructions for expert evaluation

We consider four summary models and the reference conclusion. Each model generates a radiological summary (assessment) under the specification of a source text (general examination information and radiological findings). You will be presented with the source text, the four generations from the models, and the assessment written by the physician. Please rate the generations of each model and the reference according to the following criteria: Oncological correctness, non-oncological correctness, and readability:

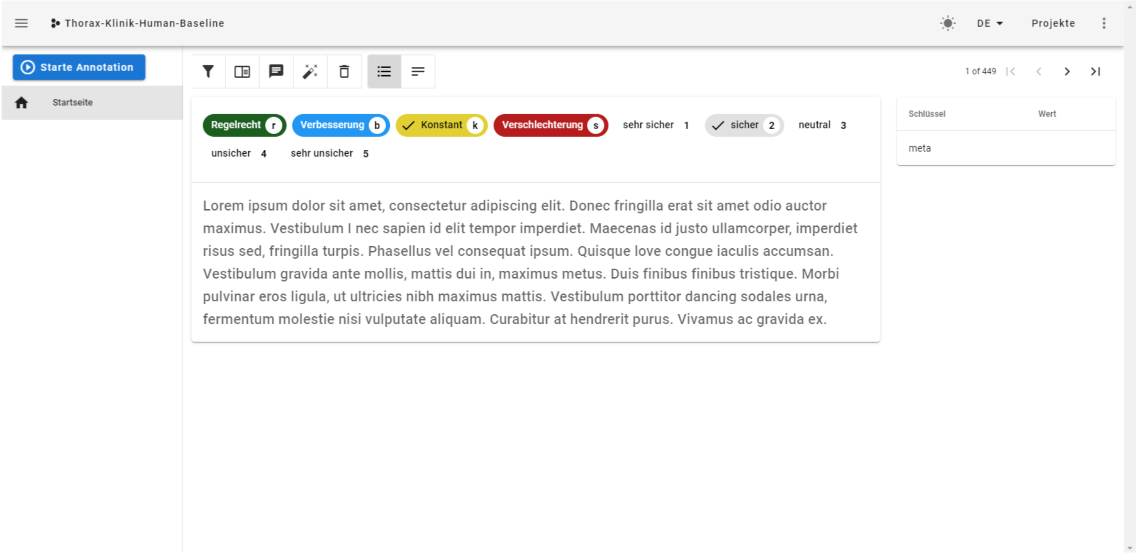
- oncological correctness: is the summary and the details about metastases (none, new, proliferation, or regressive) correct? (0) not assessable; (1) not at all correct; (2) correct to a small extent; (3) half correct; (4) correct to a large extent; (5) everything correct,
- nononcological correctness: is the general date, organ, and other information correct? (0) not assessable; (1) not at all correct; (2) correct to a small extent; (3) half correct; (4) correct to a large extent; (5) everything correct,

- readability: is the generation easy to understand, without broken expressions or unknown words? (0) not assessable; (1) many unknown words, difficult to read and comprehend; (2) several unknown words and aborted expressions, not fluent; (3) several unknown words; (4) fluent and coherent, but some unknown words; (5) correct words and expressions, fluent and coherent.

If the generation is not assessable, select 0 - not assessable. Otherwise, the scale consists of grades from 1 to 5 and must be assigned for each criterion.

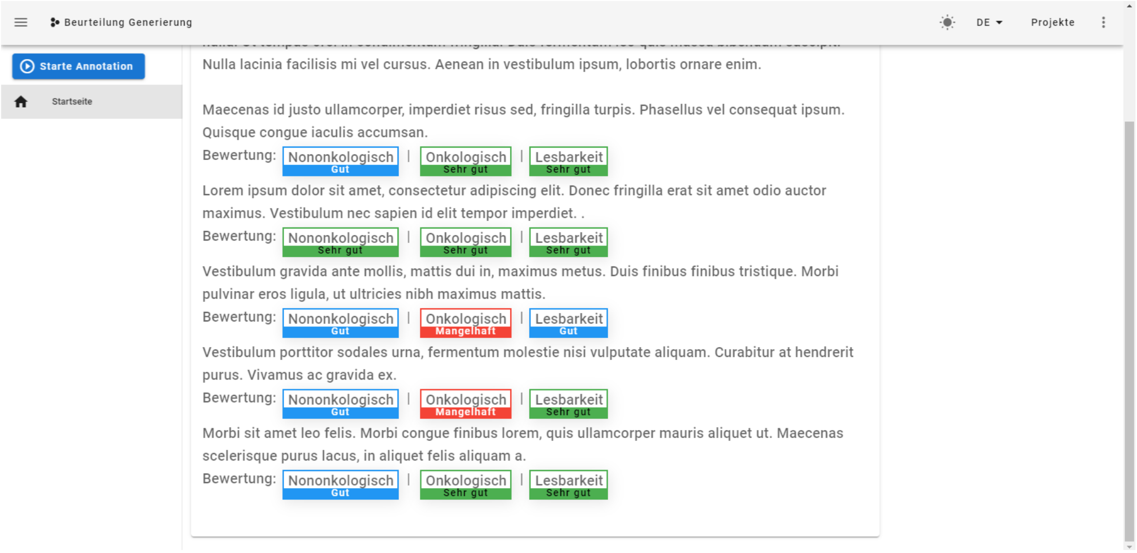
B.1.3 Screenhots of the annotation software

Figures B.1 and B.2 show screenshots of the Doccano user interface when annotating the TRC as well as when evaluating the system-generated summaries.



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec fringilla erat sit amet odio auctor maximus. Vestibulum I nec sapien id elit tempor imperdiet. Maecenas id justo ullamcorper, imperdiet risus sed, fringilla turpis. Phasellus vel consequat ipsum. Quisque love congue iaculis accumsan. Vestibulum gravida ante mollis, mattis dui in, maximus metus. Duis finibus finibus tristique. Morbi pulvinar eros ligula, ut ultricies nibh maximus mattis. Vestibulum porttitor dancing sodales urna, fermentum molestie nisi vulputate aliquam. Curabitur at hendrerit purus. Vivamus ac gravida ex.

Figure B.1: *Screenhot of the annotations of the TRC and associated certainty of the assignment. The report itself is replaced by placeholder texts due to data privacy concerns, and the labels are provided in German. The assignment is done sample by sample, while the use of keyboard shortcuts facilitates the annotation. TRC=tumor response category.*



Nulla lacinia facilisis mi vel cursus. Aenean in vestibulum ipsum, lobortis ornare enim.

Maecenas id justo ullamcorper, imperdiet risus sed, fringilla turpis. Phasellus vel consequat ipsum. Quisque congue iaculis accumsan.

Bewertung: Nononkologisch Gut | Onkologisch Sehr gut | Lesbarkeit Sehr gut

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec fringilla erat sit amet odio auctor maximus. Vestibulum nec sapien id elit tempor imperdiet. .

Bewertung: Nononkologisch Sehr gut | Onkologisch Sehr gut | Lesbarkeit Sehr gut

Vestibulum gravida ante mollis, mattis dui in, maximus metus. Duis finibus finibus tristique. Morbi pulvinar eros ligula, ut ultricies nibh maximus mattis.

Bewertung: Nononkologisch Gut | Onkologisch Mangelhaft | Lesbarkeit Gut

Vestibulum porttitor sodales urna, fermentum molestie nisi vulputate aliquam. Curabitur at hendrerit purus. Vivamus ac gravida ex.

Bewertung: Nononkologisch Gut | Onkologisch Mangelhaft | Lesbarkeit Sehr gut

Morbi sit amet leo felis. Morbi congue finibus lorem, quis ullamcorper mauris aliquet ut. Maecenas scelerisque purus lacus, in aliquet felis aliquam a.

Bewertung: Nononkologisch Gut | Onkologisch Sehr gut | Lesbarkeit Sehr gut

Figure B.2: *Screenhots of the evaluation of the system-generated conclusions. To allow the annotation of the conclusions with Doccano, a sequence labeling project is created. During the annotation, the different criteria are marked with a defined grade, similar to an entity annotation task. The report itself and the generated conclusions are replaced by placeholder texts due to data privacy concerns, and the labels are provided in German.*

B.2 Tables

Dataset Parameters	WikiLingua (#58341)	10k German News Articles Datasets (#10273)	Swiss Judgment Prediction (#45183)
Word count	405.7±242.7	365.4±270.3	398.3±264.5
Unique words	206.0±97.6	221.3±127.5	185.1±84.3
Unique bigram	362.7±210.9	345.1±249.6	301.5±178.5
Yule's I	122.9±34.8	165.3±47.9	119.4±26.8
Token type ratio	0.5±0.1	0.7±0.1	0.5±0.1
BERT split factor	1.8±0.3	1.8±0.1	1.8±0.1

Table B.1: *Results of the lexical complexity analysis for the comparison datasets. Means of the parameters are reported along with the 95% CI in parentheses. CI=Confidence Interval.*

Dataset	Annotators	Confidence	Recall (%)	Precision (%)	Accuracy (%)	F1 Score
FTOR-DKFZ	Radiologist 1	3.63±0.89	73.4 (69.9 ,77.0)	74.3 (70.6 ,78.1)	73.4 (69.9 ,77.0)	0.72 (0.69 ,0.76)
	Radiologist 2	4.22±1.05	73.6 (69.6 ,77.2)	76.0 (72.5 ,79.3)	73.6 (69.6 ,77.2)	0.74 (0.71 ,0.78)
	Student 1	4.02±1.07	72.1 (68.6 ,75.3)	73.2 (69.3 ,76.8)	72.1 (68.6 ,75.3)	0.71 (0.67 ,0.75)
	Student 2	3.13±0.79	64.8 (61.2 ,68.6)	64.0 (59.4 ,68.5)	64.8 (61.2 ,68.6)	0.62 (0.58 ,0.66)
	RT 1	3.20±0.98	66.9 (63.7 ,69.9)	70.1 (64.1 ,74.8)	66.9 (63.7 ,69.9)	0.62 (0.59 ,0.65)
	RT 2	3.23±1.00	59.9 (56.1 ,63.7)	64.4 (59.9 ,68.9)	59.9 (56.1 ,63.7)	0.59 (0.55 ,0.63)
	RT 3	2.16±0.96	48.0 (44.4 ,51.5)	63.8 (60.7 ,66.9)	48.0 (44.4 ,51.5)	0.45 (0.41 ,0.49)
	Radiologists (MiA)	3.92±1.02	73.5 (70.5 ,76.3)	74.2 (71.3 ,77.0)	73.5 (70.5 ,76.3)	0.74 (0.71 ,0.76)
	Students (MiA)	3.58±1.04	68.5 (65.7 ,71.3)	68.8 (65.6 ,71.9)	68.5 (65.7 ,71.3)	0.67 (0.64 ,0.70)
	RTs (MiA)	2.86±1.10	58.3 (55.8 ,60.7)	62.8 (59.4 ,66.1)	58.3 (55.8 ,60.7)	0.56 (0.53 ,0.58)
FTORT-TKH	Radiologist 1	3.66±0.75	86.6 (84.1 ,89.1)	86.1 (83.3 ,88.9)	86.6 (84.1 ,89.1)	0.86 (0.83 ,0.89)
	Radiologist 2	4.36±0.96	81.9 (79.0 ,85.0)	86.0 (83.7 ,88.0)	81.9 (79.0 ,85.0)	0.83 (0.80 ,0.85)
	Student 1	3.91±0.87	87.3 (84.8 ,89.6)	86.2 (83.6 ,88.5)	87.3 (84.8 ,89.6)	0.87 (0.84 ,0.89)
	Student 2	3.32±0.70	71.4 (68.1 ,74.6)	81.0 (78.7 ,83.0)	71.4 (68.1 ,74.6)	0.72 (0.69 ,0.75)
	RT 1	3.15±0.86	84.1 (81.3 ,86.8)	82.9 (80.2 ,85.6)	84.1 (81.3 ,86.8)	0.83 (0.81 ,0.86)
	RT 2	3.23±0.92	78.3 (75.3 ,81.3)	79.7 (76.6 ,82.7)	78.3 (75.3 ,81.3)	0.79 (0.76 ,0.82)
	RT 3	2.17±1.03	62.6 (59.4 ,65.8)	66.4 (62.9 ,70.0)	62.6 (59.4 ,65.8)	0.60 (0.56 ,0.64)
	Radiologists (MiA)	4.01±0.93	84.3 (82.2 ,86.3)	85.0 (83.1 ,86.8)	84.3 (82.2 ,86.3)	0.84 (0.82 ,0.86)
	Students (MiA)	3.61±0.84	79.3 (77.1 ,81.4)	81.6 (79.7 ,83.4)	79.3 (77.1 ,81.4)	0.79 (0.77 ,0.81)
	RTs (MiA)	2.85±1.06	75.0 (73.0 ,77.1)	75.3 (73.1 ,77.3)	75.0 (73.0 ,77.1)	0.74 (0.72 ,0.76)
FTORs (MiA)	Radiologist 1	3.64±0.82	80.5 (78.4 ,82.7)	80.3 (78.0 ,82.6)	80.5 (78.4 ,82.7)	0.80 (0.78 ,0.82)
	Radiologist 2	4.29±1.01	78.1 (75.6 ,80.5)	80.2 (78.1 ,82.3)	78.1 (75.6 ,80.5)	0.79 (0.76 ,0.81)
	Student 1	3.96±0.97	80.3 (78.2 ,82.3)	79.7 (77.4 ,81.9)	80.3 (78.2 ,82.3)	0.79 (0.77 ,0.81)
	Student 2	3.24±0.75	68.4 (65.8 ,70.8)	70.4 (67.6 ,73.0)	68.4 (65.8 ,70.8)	0.67 (0.65 ,0.70)
	RT 1	3.18±0.92	76.2 (74.2 ,78.2)	77.0 (73.7 ,79.7)	76.2 (74.2 ,78.2)	0.73 (0.71 ,0.76)
	RT 2	3.23±0.96	69.8 (67.5 ,72.3)	69.1 (66.6 ,71.8)	69.8 (67.5 ,72.3)	0.68 (0.66 ,0.71)
	RT 3	2.16±1.00	55.9 (53.5 ,58.4)	63.7 (61.1 ,66.3)	55.9 (53.5 ,58.4)	0.52 (0.49 ,0.55)
	Radiologists (MiA)	3.97±0.97	79.3 (77.5 ,81.0)	79.7 (78.0 ,81.4)	79.3 (77.5 ,81.0)	0.79 (0.78 ,0.81)
	Students (MiA)	3.60±0.94	74.3 (72.6 ,76.1)	74.5 (72.5 ,76.4)	74.3 (72.6 ,76.1)	0.73 (0.72 ,0.75)
	RTs (MiA)	2.86±1.08	67.3 (65.7 ,68.8)	66.9 (64.9 ,68.8)	67.3 (65.7 ,68.8)	0.65 (0.63 ,0.67)

Table B.2: *TRC classification results, micro-averaged, for the human annotators and the and the three human annotator groups, respectively, on the FTOR-DKFZ and FTORT-TKH and across all FTORs. Reported scores are mean values with 95% CIs in parenthesis unless otherwise noted. The confidences assigned during annotations are reported as means ± STDs. CI=Confidence Interval, FTOR=free-text-oncology report, FTOR-DKFZ=free-text-oncology reports from the German Cancer Research Center, FTORT-TKH=free-text-oncology reports from the Heidelberg Thoracic Clinic, MiA=micro-averaged (Calculation of metrics over a concatenated list of labels and predictions.), STD=standard deviation, TRC=tumor response category.*

Table B.3: *List of cases used for testing in the experiments on the seen data from the pre-processed multi-site prostate MRI segmentation dataset (liuquande.github.io/SAML/). This table was adapted from Kades et al. (2022b) with permission of Springer. MRI=magnetic resonance imaging.*

RUNMC	Case24	Case12	Case05	Case07	Case08	Case26	Case10	Case17	Case01
BMC	Case24	Case12	Case05	Case07	Case08	Case26	Case10	Case17	Case01
I2CVB	Case07	Case00	Case10	Case05	Case15	Case12			
UCL	Case26	Case30	Case33	Case36					
BIDMC	Case06	Case05	Case08	Case07					
HK	Case38	Case41	Case42	Case40					

Table B.4: *Mean DICE (%) scores with standard deviation for each dataset, algorithm, architecture, and experiment. This table was adapted from Kades et al. (2022b) with permission of Springer. DICE=Sørensen-Dice coefficient.*

	Dataset	RUNMC	BMC	I2CVB	UCL	BIDMC	HK	Average
Setup	Algorithm	# 30	# 30	# 19	# 13	# 12	# 12	
Seen								
2D	Intra-site	87.74 (3.28)	91.14 (2.40)	81.12 (5.45)	88.06 (1.87)	69.83 (16.88)	85.11 (6.84)	83.83 (7.64)
	DeepAll	88.55 (2.68)	91.04 (2.82)	79.21 (9.39)	90.14 (1.77)	80.98 (11.16)	89.46 (3.57)	86.57 (5.11)
	Federated	88.27 (2.65)	90.88 (3.01)	84.50 (3.08)	90.59 (1.72)	78.01 (16.36)	88.97 (2.86)	86.87 (4.91)
3D	RUNMC	89.58 (3.34)	68.44 (22.21)	22.41 (31.12)	82.92 (4.35)	46.72 (28.38)	81.88 (10.26)	65.33 (25.96)
	BMC	46.20 (24.16)	90.46 (1.84)	16.11 (16.70)	64.13 (19.16)	41.72 (32.62)	40.56 (24.37)	49.86 (25.13)
	I2CVB	55.87 (14.02)	69.81 (10.02)	83.64 (6.02)	28.93 (23.12)	26.38 (25.02)	20.53 (22.44)	47.53 (26.05)
	UCL	85.76 (2.37)	75.70 (9.37)	55.13 (14.80)	88.19 (3.83)	38.61 (22.13)	56.40 (20.36)	66.63 (19.67)
	BIDMC	40.11 (14.92)	78.01 (17.44)	10.40 (9.08)	29.66 (11.19)	73.95 (29.08)	47.87 (35.49)	46.67 (25.98)
	HK	80.32 (4.26)	47.39 (27.64)	19.58 (29.58)	70.03 (17.59)	54.22 (26.59)	84.96 (5.84)	59.42 (24.33)
	Ensemble	87.48 (2.85)	86.27 (4.71)	48.28 (23.93)	88.02 (3.51)	58.32 (26.28)	82.51 (7.49)	75.15 (17.32)
	DeepAll	90.00 (2.13)	91.57 (1.98)	82.27 (7.07)	90.02 (1.61)	87.64 (4.72)	90.49 (2.75)	88.66 (3.38)
	Federated	89.96 (1.96)	91.50 (1.48)	84.50 (5.12)	90.16 (2.96)	87.70 (3.18)	90.99 (2.83)	89.14 (2.62)
Unseen								
2D	DeepAll	84.89 (4.37)	83.10 (4.75)	71.17 (16.77)	85.88 (4.44)	74.18 (12.20)	86.24 (4.95)	80.91 (6.54)
	Federated	85.84 (3.93)	81.96 (5.73)	76.52 (9.94)	84.94 (4.24)	73.19 (15.35)	86.09 (4.30)	81.42 (5.40)
3D	RUNMC		70.06 (18.61)	29.73 (25.33)	82.03 (4.59)	59.81 (22.39)	86.60 (3.31)	65.65 (22.64)
	BMC	48.30 (25.55)		17.78 (12.18)	62.60 (18.68)	58.27 (24.21)	36.67 (20.12)	44.73 (18.07)
	I2CVB	45.94 (18.21)	54.89 (26.19)		33.49 (12.75)	19.34 (25.01)	40.00 (18.58)	38.73 (13.40)
	UCL	82.35 (10.23)	79.21 (12.34)	49.16 (18.35)		59.51 (17.40)	74.71 (15.95)	68.99 (14.13)
	BIDMC	37.15 (17.95)	80.74 (12.26)	17.18 (12.88)	37.44 (17.92)		59.98 (25.41)	46.50 (24.41)
	HK	79.66 (16.64)	42.76 (25.88)	34.14 (21.27)	61.51 (17.04)	56.26 (18.91)		54.86 (17.58)
	Ensemble	76.53 (16.13)	84.99 (4.69)	49.14 (20.15)	84.34 (7.26)	72.15 (13.57)	85.81 (8.39)	75.49 (14.01)
	DeepAll	83.97 (10.53)	80.37 (15.68)	58.45 (18.01)	85.59 (5.60)	78.98 (13.88)	89.24 (3.80)	79.43 (10.92)
	Federated	85.01 (6.95)	85.36 (8.29)	67.63 (13.55)	86.97 (4.58)	81.95 (9.62)	88.51 (4.48)	82.57 (7.64)

Table B.5: *Mean ASD (mm) scores with standard deviation for each dataset, algorithm, architecture, and experiment. This table was adapted from Kades et al. (2022b) with permission of Springer. ASD=Average Surface Distance.*

	Dataset	RUNMC	BMC	I2CVB	UCL	BIDMC	HK	Average
Setup	Algorithm	# 30	# 30	# 19	# 13	# 12	# 12	
Seen								
2D	Intra-site	0.79 (0.27)	0.72 (0.20)	2.05 (0.42)	0.82 (0.21)	2.35 (1.45)	1.08 (0.56)	1.30 (0.71)
	DeepAll	0.73 (0.15)	0.73 (0.24)	2.32 (0.95)	0.67 (0.18)	1.58 (0.61)	0.71 (0.17)	1.12 (0.68)
	Federated	0.77 (0.14)	0.70 (0.26)	2.00 (0.50)	0.61 (0.15)	1.62 (0.72)	0.77 (0.21)	1.08 (0.58)
3D	RUNMC	0.78 (0.48)	26.53 (24.91)	55.57 (34.37)	2.65 (1.41)	50.49 (40.48)	8.74 (15.02)	24.13 (24.22)
	BMC	81.22 (39.26)	0.74 (0.23)	105.08 (16.20)	52.26 (40.98)	49.59 (32.75)	84.49 (32.72)	62.23 (36.69)
	I2CVB	43.17 (28.69)	20.66 (11.11)	2.14 (0.82)	99.85 (34.76)	28.10 (27.37)	23.00 (23.65)	36.16 (33.89)
	UCL	0.97 (0.23)	20.92 (19.70)	27.24 (15.82)	1.25 (1.19)	41.54 (19.57)	24.34 (19.53)	19.38 (15.80)
	BIDMC	96.77 (15.00)	17.43 (27.61)	108.40 (12.31)	96.33 (22.95)	40.65 (78.91)	70.55 (37.36)	71.69 (36.03)
	HK	3.71 (3.91)	51.85 (43.15)	59.50 (43.95)	8.51 (11.65)	27.14 (19.08)	1.01 (0.49)	25.29 (25.36)
	Ensemble	0.92 (0.43)	3.46 (7.80)	20.93 (18.27)	0.88 (0.37)	15.54 (14.82)	8.00 (13.92)	8.29 (8.31)
	DeepAll	0.67 (0.17)	0.64 (0.20)	2.14 (0.64)	0.70 (0.18)	1.26 (0.13)	0.66 (0.21)	1.01 (0.60)
	Federated	0.69 (0.11)	0.61 (0.14)	1.95 (0.52)	0.63 (0.21)	1.28 (0.19)	0.62 (0.20)	0.96 (0.55)
Unseen								
2D	DeepAll	1.37 (0.88)	1.26 (0.36)	4.54 (2.18)	1.04 (0.48)	4.73 (4.73)	1.20 (0.73)	2.36 (1.77)
	Federated	1.11 (0.41)	1.33 (0.41)	4.52 (2.94)	1.53 (1.80)	2.56 (1.66)	1.03 (0.37)	2.01 (1.35)
3D	RUNMC		23.80 (26.59)	63.48 (44.65)	1.77 (0.71)	35.85 (19.79)	1.03 (0.32)	25.19 (26.05)
	BMC	74.24 (42.88)		103.77 (10.64)	56.90 (40.17)	39.14 (25.16)	83.32 (30.79)	71.48 (24.74)
	I2CVB	81.40 (48.01)	60.58 (38.10)		107.41 (18.45)	24.05 (24.31)	24.90 (22.27)	59.67 (36.16)
	UCL	4.25 (16.68)	14.51 (21.16)	32.33 (23.76)		35.64 (19.79)	11.66 (19.13)	19.68 (13.64)
	BIDMC	101.29 (24.97)	18.94 (37.71)	111.46 (18.13)	99.51 (26.58)		65.02 (42.30)	79.24 (37.99)
	HK	12.33 (26.34)	58.34 (34.11)	49.64 (22.18)	33.41 (18.92)	35.39 (23.67)		37.82 (17.58)
	Ensemble	38.57 (41.58)	2.25 (4.77)	37.49 (23.64)	16.68 (30.25)	18.96 (15.73)	5.72 (14.69)	19.95 (15.37)
	DeepAll	4.91 (17.82)	16.77 (30.06)	24.77 (29.58)	8.34 (18.48)	25.48 (33.67)	1.47 (1.76)	13.62 (10.26)
	Federated	3.65 (14.10)	8.05 (17.43)	16.34 (16.28)	1.78 (2.73)	21.16 (26.68)	1.86 (3.77)	8.81 (8.18)

Bibliography

Nicole Agaronnik, Charlotta Lindvall, Areej El-Jawahri, Wei He, and Lisa Iezzoni. **Use of Natural Language Processing to Assess Frequency of Functional Status Documentation for Patients Newly Diagnosed With Colorectal Cancer**. *JAMA Oncology*, 6(10):1628–1630, 10 2020. ISSN 2374-2437. doi: 10.1001/jamaoncol.2020.2708. URL `https://doi.org/10.1001/jamaoncol.2020.2708`.

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. **ETC: Encoding Long and Structured Inputs in Transformers**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.19. URL `https://aclanthology.org/2020.emnlp-main.19`.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. **Optuna: A Next-generation Hyperparameter Optimization Framework**. *arXiv e-prints*, art. arXiv:1907.10902, July 2019.

João Rafael Almeida, Eriksson Monteiro, and José Luís Oliveira. **An Architecture to Define Cohorts over Medical Imaging Datasets**. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 545–549, 2021. doi: 10.1109/CBMS52027.2021.00088.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. **Publicly Available Clinical BERT Embeddings**. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://aclanthology.org/W19-1909`.

Amazon. **Amazon Simple Storage Service S3**. https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html, 2023a. Accessed: January 18, 2023.

Amazon. **Amazon Web Services (AWS) - Cloud Computing Services**. https://aws.amazon.com/, 2023b. Accessed: January 18, 2023.

Andrey Fedorov. **QIICR**. http://qiicr.org/index.html, 2023. Accessed: January 18, 2023.

Apache. **Apache Airflow**. https://airflow.apache.org/, 2023. Accessed: January 18, 2023.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. **"What is relevant in a text document?": An interpretable machine learning approach**. *PLOS ONE*, 12(8):1–23, 08 2017. doi: 10.1371/journal.pone.0181142. URL `https://doi.org/10.1371/journal.pone.0181142`.

Ron Artstein and Massimo Poesio. **Survey Article: Inter-Coder Agreement for Computational Linguistics**. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL `https://aclanthology.org/J08-4004`.

Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin, Allison W. Kurian, and Daniel L. Rubin. **Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer**. *JCO Clinical Cancer Informatics*, (3): 1–12, 2019. doi: 10.1200/CCI.19.00034. URL `https://doi.org/10.1200/CCI.19.00034`. PMID: 31584836.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. **Longformer: The Long-Document Transformer**. *arXiv e-prints*, art. arXiv:2004.05150, April 2020.

BioCreative/OHNLP Challenge 2018. **BioCreative/OHNLP Challenge 2018**. https://sites.google.com/view/ohnlp2018/home, 2023. Accessed: January 20, 2023.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. **Communication-Efficient Learning of Deep Networks from Decentralized Data**. *arXiv e-prints*, art. arXiv:1602.05629, February 2016.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. **Faithful to the Original: Fact-Aware Neural Abstractive Summarization**. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. **SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation**. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics,

2017. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`.

David Chang, Eric Lin, Cynthia Brandt, and Richard Andrew Taylor. **Incorporating Domain Knowledge Into Language Models by Using Graph Convolutional Networks for Assessing Semantic Textual Similarity: Model Development and Performance Comparison**. *JMIR Med Inform*, 9(11):e23101, Nov 2021. ISSN 2291-9694. doi: 10.2196/23101. URL `https://medinform.jmir.org/2021/11/e23101`.

Kushal Chawla, Kundan Krishna, and Balaji Vasan Srinivasan. **Improving generation quality of pointer networks via guided attention**. *arXiv e-prints*, art. arXiv:1901.11492, Jan 2019.

David Chen, Sijia Liu, Paul Kingsbury, Sunghwan Sohn, Curtis B. Storlie, Elizabeth B. Habermann, James M. Naessens, David W. Larson, and Hongfang Liu. **Deep learning and alternative learning strategies for retrospective real-world clinical data**. *npj Digital Medicine*, 2(1):43, May 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0122-0. URL `https://doi.org/10.1038/s41746-019-0122-0`.

Qingyu Chen, Jingcheng Du, Sun Kim, W. Wilbur, and Zhiyong lu. **Combining rich features and deep learning for finding similar sentences in electronic medical records**. 2018.

Yen-Chun Chen and Mohit Bansal. **Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1063. URL `https://www.aclweb.org/anthology/P18-1063`.

Sumit Chopra, Michael Auli, and Alexander M. Rush. **Abstractive Sentence Summarization with Attentive Recurrent Neural Networks**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1012. URL `https://aclanthology.org/N16-1012`.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. **The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository**. *Journal of Digital Imaging*, 26(6): 1045–1057, Dec 2013. ISSN 1618-727X. doi: 10.1007/s10278-013-9622-7. URL `https://doi.org/10.1007/s10278-013-9622-7`.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. **ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators**. In *ICLR*, 2020. URL `https://openreview.net/pdf?id=r1xMH1BtvB`.

Serverless Working Group CNCF. **CNCF WG-Serverless Whitepaper v1.0**. https://github.com/cncf/wg-serverless, 2023. Accessed: January 18, 2023.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. **Supervised Learning of Universal Sentence Representations from Natural Language Inference Data**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL `https://aclanthology.org/D17-1070`.

Christine M. Cutillo, Karlie R. Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D. Mandl, Tyler Beck, Elaine Collier, Christine Colvis, Kenneth Gersing, Valery Gordon, Roxanne Jensen, Behrouz Shabestari, Noel Southall, and M. I. in Healthcare Workshop Working Group. **Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency**. *npj Digital Medicine*, 3(1):47, Mar 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0254-2. URL `https://doi.org/10.1038/s41746-020-0254-2`.

Bijoyan Das and Sarit Chakraborty. **An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation**. *arXiv e-prints*, art. arXiv:1806.06407, June 2018.

Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. **Big Data in Healthcare: Management, Analysis and Future Prospects**. *Journal of Big Data*, 6(1):54, June 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0217-0.

Thomas Davenport and Ravi Kalakota. **The potential for artificial intelligence in healthcare**. *Future Healthcare Journal*, 6(2):94–98, 2019. ISSN 2514-6645. doi: 10.7861/futurehosp.6-2-94. URL `https://www.rcpjournals.org/content/6/2/94`.

Ittai Dayan, Holger R. Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z. Abidin, Andrew Liu, Anthony Beardsworth Costa, Bradford J. Wood, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, C. K. Lee, Peiying Ruan, Daguang Xu, Dufan Wu, Eddie Huang, Felipe Campos Kitamura, Griffin Lacey, Gustavo César de Antônio Corradi, Gustavo Nino, Hao-Hsin Shin, Hirofumi Obinata, Hui Ren, Jason C. Crane, Jesse Tetreault, Jiahui Guan, John W. Garrett, Joshua D. Kaggie, Jung Gil Park, Keith Dreyer, Krishna Juluru, Kristopher Kersten, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena Abdel-Maseeh, Nicola Rieke, Pablo F. Damasceno, Pedro Mario Cruz e Silva, Pochuan

Wang, Sheng Xu, Shuichi Kawano, Sira Sriswasdi, Soo Young Park, Thomas M. Grist, Varun Buch, Watsamon Jantarabenjakul, Weichung Wang, Won Young Tak, Xiang Li, Xihong Lin, Young Joon Kwon, Abood Quraini, Andrew Feng, Andrew N. Priest, Baris Turkbey, Benjamin Glicksberg, Bernardo Bizzo, Byung Seok Kim, Carlos Tor-Díez, Chia-Cheng Lee, Chia-Jung Hsu, Chin Lin, Chiu-Ling Lai, Christopher P. Hess, Colin Compas, Deepeksha Bhatia, Eric K. Oermann, Evan Leibovitz, Hisashi Sasaki, Hitoshi Mori, Isaac Yang, Jae Ho Sohn, Krishna Nand Keshava Murthy, Li-Chen Fu, Matheus Ribeiro Furtado de Mendonça, Mike Fralick, Min Kyu Kang, Mohammad Adil, Natalie Gangai, Peerapon Vateekul, Pierre Elnajjar, Sarah Hickman, Sharmila Majumdar, Shelley L. McLeod, Sheridan Reed, Stefan Gräf, Stephanie Harmon, Tatsuya Kodama, Thanyawee Puthanakit, Tony Mazzulli, Vitor Lima de Lavor, Yothin Rakvongthai, Yu Rim Lee, Yuhong Wen, Fiona J. Gilbert, Mona G. Flores, and Quanzheng Li. **Federated learning for predicting clinical outcomes in patients with COVID-19**. *Nature Medicine*, 27(10): 1735–1743, Oct 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01506-3. URL `https://doi.org/10.1038/s41591-021-01506-3`.

DBMI Data Portal. **n2c2 NLP Research Data Sets**. https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp, 2023. Accessed: January 20, 2023.

dcmjs. **dcmjs**. https://github.com/commontk/dcmjs, 2023. Accessed: January 20, 2023.

Tom De Smedt and Walter Daelemans. **Pattern for Python**. *J. Mach. Learn. Res.*, 13 (null):20632067, jun 2012. ISSN 1532-4435.

John Deaton. **Transformers and Pointer-Generator Networks for Abstractive Summarization**. 2019.

deepset. **deepset**. https://www.deepset.ai/german-bert, 2023. Accessed: January 20, 2023.

Shrey Desai and Greg Durrett. **Calibration of Pre-trained Transformers**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL `https://aclanthology.org/2020.emnlp-main.21`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics,

2019. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Oliver Diaz, Kaisar Kushibar, Richard Osuala, Akis Linardos, Lidia Garrucho, Laura Igual, Petia Radeva, Fred Prior, Polyxeni Gkontra, and Karim Lekadir. **Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools**. *Physica Medica*, 83:25–37, 2021. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2021.02.007. URL `https://www.sciencedirect.com/science/article/pii/S1120179721000958`.

Django. **Django**. https://www.djangoproject.com/, 2023. Accessed: January 20, 2023.

Qifei Dong, Gang Luo, David Haynor, Michael OReilly, Ken Linnau, Ziv Yaniv, Jeffrey G. Jarvik, and Nathan Cross. **DicomAnnotator: a Configurable Open-Source Software Program for Efficient DICOM Image Annotation**. *J Digit Imaging*, 33 (6):1514–1526, 2020. ISSN 1618-727X. doi: 10.1007/s10278-020-00370-w. URL `https://doi.org/10.1007/s10278-020-00370-w`.

Qi Dou, Tiffany Y. So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather H. C. Lee, Kevin Yu, Zuxin Feng, Li Dong, Egon Burian, Friederike Jungmann, Rickmer Braren, Marcus Makowski, Bernhard Kainz, Daniel Rueckert, Ben Glocker, Simon C. H. Yu, and Pheng Ann Heng. **Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study**. *npj Digital Medicine*, 4(1):60, Mar 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00431-6. URL `https://doi.org/10.1038/s41746-021-00431-6`.

Bradley Efron. **Second Thoughts on the Bootstrap**. *Statistical Science*, 18(2):135 – 140, 2003. doi: 10.1214/ss/1063994968. URL `https://doi.org/10.1214/ss/1063994968`.

Bradley Efron, Robert Tibshirani, and R J Tibshirani. **An introduction to the bootstrap**. Chapman & Hall/CRC, Philadelphia, PA, 1994.

E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. **New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)**. *European Journal of Cancer*, 45(2): 228–247, 2009. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2008.10.026. URL `https://www.sciencedirect.com/science/article/pii/S0959804908008733`. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers.

Peter J. Embi, Charlene Weir, Efthimis N. Efthimiadis, Stephen M. Thielke, Ashley N. Hedeen, and Kenric W. Hammond. **Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators**. 20(4):718–726, 2013. ISSN 1067-5027. doi: 10.1136/amiajnl-2012-000946. URL `https://academic.oup.com/jamia/article/20/4/718/2909346`.

European ESR. **ESR paper on structured reporting in radiology**. *Insights into Imaging*, 9(1):1–7, 02 2018. ISSN 1869-4101. doi: 10.1007/s13244-017-0588-8. URL `https://doi.org/10.1007/s13244-017-0588-8`.

FAIR Principles. **FAIR Principles**. https://www.go-fair.org/fair-principles/, 2023. Accessed: January 20, 2023.

Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. **WikiLingua: A New Benchmark Dataset for Multilingual Abstractive Summarization**. In *Findings of EMNLP, 2020*, 2020.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. **Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213. URL `https://aclanthology.org/P19-1213`.

Hongjie Fan, Zhiyi Ma, Hongqiang Li, Dongsheng Wang, and Junfei Liu. **Enhanced answer selection in CQA using multi-dimensional features combination**. 24 (3):346–359, 2019. ISSN 1007-0214. doi: 10.26599/TST.2018.9010050.

FATE community. **FATE**. https://fate.fedai.org/, 2023. Accessed: January 20, 2023.

FedML. **FedML**. https://fedml.ai/, 2023. Accessed: January 20, 2023.

Ines Feki, Sourour Ammar, Yousri Kessentini, and Khan Muhammad. **Federated learning for COVID-19 screening from Chest X-ray images**. *Applied Soft Computing*, 106:107330, 2021. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107330. URL `https://www.sciencedirect.com/science/article/pii/S1568494621002532`.

Matthias A. Fink, Klaus Kades, Arved Bischoff, Martin Moll, Merle Schnell, Maike Küchler, Gregor Köhler, Jan Sellner, Claus Peter Heussel, Hans-Ulrich Kauczor, Heinz-Peter Schlemmer, Klaus Maier-Hein, Tim F. Weber, and Jens Kleesiek. **Deep Learningbased Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports**. *Radiology: Artificial Intelligence*, 4

(5):e220055, 2022a. doi: 10.1148/ryai.220055. URL `https://doi.org/10.1148/ryai.220055`.

Matthias A. Fink, Victoria L. Mayer, Thomas Schneider, Constantin Seibold, Rainer Stiefelhagen, Jens Kleesiek, Tim F. Weber, and Hans-Ulrich Kauczor. **CT Angiography Clot Burden Score from Data Mining of Structured Reports for Pulmonary Embolism**. *Radiology*, 302(1):175–184, 2022b. doi: 10.1148/radiol.2021211013. URL `https://doi.org/10.1148/radiol.2021211013`. PMID: 34581626.

Fraunhofer MEVIS. **Satori**. https://www.mevis.fraunhofer.de/en/research-and-technologies/ai-collaboration-toolkit.html, 2023. Accessed: January 20, 2023.

Peter M. Full, Fabian Isensee, Paul F. Jäger, and Klaus Maier-Hein. **Studying Robustness of Semantic Segmentation Under Domain Shift in Cardiac MRI**. In Esther Puyol Anton, Mihaela Pop, Maxime Sermesant, Victor Campello, Alain Lalande, Karim Lekadir, Avan Suinesiaputra, Oscar Camara, and Alistair Young, editors, *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 238–249, Cham, 2021. Springer International Publishing. ISBN 978-3-030-68107-4.

Romane Gauriau, Christopher Bridge, Lina Chen, Felipe Kitamura, Neil A. Tenenholtz, John E. Kirsch, Katherine P. Andriole, Mark H. Michalski, and Bernardo C. Bizzo. **Using DICOM Metadata for Radiological Image Series Categorization: a Feasibility Study on Large Clinical Brain MRI Datasets**. *J Digit Imaging*, 33 (3):747–762, 2020. ISSN 1618-727X. doi: 10.1007/s10278-019-00308-x. URL `https://doi.org/10.1007/s10278-019-00308-x`.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T. Carlson, Joy T. Wu, Jonathan Welt, John Foote, Jr., Edward T. Moseley, David W. Grant, Patrick D. Tyler, and Leo A. Celi. **Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives**. *PLOS ONE*, 13(2):1–19, 02 2018. doi: 10.1371/journal.pone.0192360. URL `https://doi.org/10.1371/journal.pone.0192360`.

Gender Guesser. **Gender Guesser**. https://github.com/lead-ratings/gender-guesser, 2023. Accessed: January 20, 2023.

Brad W. Genereaux, Donald K. Dennison, Kinson Ho, Robert Horn, Elliot Lewis Silver, Kevin O'Donnell, and Charles E. Kahn. **DICOMweb™: Background and Application of the Web Standard for Medical Imaging**. *Journal of Digital Imaging*, 31 (3):321–326, June 2018. ISSN 1618-727X. doi: 10.1007/s10278-018-0073-z.

Richard F. Gillum. **From Papyrus to the Electronic Tablet: A Brief History of the Clinical Medical Record with Lessons for the Digital Age**. *The American Journal of Medicine*, 126(10):853–857, 2013. ISSN 0002-9343. doi: https://doi.org/10.1016/j.amjmed.2013.03.024. URL `https://www.sciencedirect.com/science/article/pii/S0002934313003987`.

GloVe. **GloVe**. https://github.com/stanfordnlp/GloVe, 2023. Accessed: January 20, 2023.

GLUE Benchmark. **GLUE Benchmark**. https://gluebenchmark.com/, 2023. Accessed: January 20, 2023.

Wael Gomaa and Aly Fahmy. **A Survey of Text Similarity Approaches**. 68, 2013. doi: 10.5120/11638-7118.

Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. **Detecting When Pre-trained nnU-Net Models Fail Silently for Covid-19 Lung Lesion Segmentation**, pages 304–314. 09 2021. ISBN 978-3-030-87233-5. doi: 10.1007/978-3-030-87234-2_29.

Grafana. **Grafana**. https://grafana.com/, 2023. Accessed: January 20, 2023.

Ran Gu, Jingyang Zhang, Rui Huang, Wenhui Lei, Guotai Wang, and Shaoting Zhang. **Domain Composition and Attention forăUnseen-Domain Generalizable Medical Image Segmentation**. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 241–250, Cham, 2021a. Springer International Publishing. ISBN 978-3-030-87199-4.

Ran Gu, Jingyang Zhang, Rui Huang, Wenhui Lei, Guotai Wang, and Shaoting Zhang. **Domain Composition and Attention forăUnseen-Domain Generalizable Medical Image Segmentation**. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 241–250. Springer, 2021b. ISBN 978-3-030-87199-4.

Gunter Zeilinger. **Open Source Clinical Image and Object Management**. https://dcm4che.org/, October 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. **On Calibration of Modern Neural Networks**. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. **Self-Attention Attribution: Interpreting Information Interactions Inside Transformer**. *arXiv e-prints*, art. arXiv:2004.11207, April 2020.

Sadid A. Hasan and Oladimeji Farri. **Clinical Natural Language Processing with Deep Learning**, pages 147–171. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05249-2. doi: 10.1007/978-3-030-05249-2_5. URL `https://doi.org/10.1007/978-3-030-05249-2_5`.

Geoffrey E Hinton and Sam Roweis. **Stochastic Neighbor Embedding**. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL `https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf`.

HL7. **FHIR v4.0.1**, October 2019.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. **A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1013. URL `https://aclanthology.org/P18-1013`.

Kexin Huang, Sankeerth Garapati, and Alexander S. Rich. **An Interpretable End-to-end Fine-tuning Approach for Long Clinical Text**. *arXiv e-prints*, art. arXiv:2011.06504, November 2020.

Peng Huang, Cheng T. Lin, Yuliang Li, Martin C. Tammemagi, Malcolm V. Brock, Sukhinder Atkar-Khattra, Yanxun Xu, Ping Hu, John R. Mayo, Heidi Schmidt, Michel Gingras, Sergio Pasian, Lori Stewart, Scott Tsai, Jean M. Seely, Daria Manos, Paul Burrowes, Rick Bhatia, Ming-Sound Tsao, and Stephen Lam. **Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method**. *The Lancet Digital Health*, 1(7):e353–e362, Nov 2019. ISSN 2589-7500. doi: 10.1016/S2589-7500(19)30159-1. URL `https://doi.org/10.1016/S2589-7500(19)30159-1`.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. **Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines**. *npj Digital Medicine*, 3(1):136, Oct 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z. URL `https://doi.org/10.1038/s41746-020-00341-z`.

Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. **An Efficient Approach for Assessing Hyperparameter Importance**. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 754–762, Bejing, China, 22–24 Jun 2014. PMLR. URL `https://proceedings.mlr.press/v32/hutter14.html`.

Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S. Bitterman, Steven F. Petit, Daphne A. Haas-Kogan, Benjamin Kann, Hugo J. W. L. Aerts, and Raymond H. Mak. **Artificial intelligence in radiation oncology**. *Nature Reviews Clinical Oncology*, 17(12):771–781, Dec 2020. ISSN 1759-4782. doi: 10.1038/s41571-020-0417-8. URL `https://doi.org/10.1038/s41571-020-0417-8`.

I2CVB. **Initiative for Collaborative Computer Vision Benchmarking**. https://i2cvb.github.io/, 2023. Accessed: January 20, 2023.

Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. **Uncertainty Estimates and Multi-hypotheses Networks for Optical Flow**. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 677–693, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

The Neuroimaging Informatics Technology Initiative. **NIfTI: — Neuroimaging Informatics Technology Initiative**. https://nifti.nimh.nih.gov/, March 2011.

Fabian Isensee, Paul Jäger, Jakob Wasserthal, David Zimmerer, Jens Petersen, Simon Kohl, Justus Schock, Andre Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. **Batchgenerators - a Python Framework for Data Augmentation**. Zenodo, Jan 2020.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**. *Nature Methods*, 18(2):203–211, Feb 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL `https://doi.org/10.1038/s41592-020-01008-z`.

Paul F. Jäger, Sebastian Bickelhaupt, Frederik Bernd Laun, Wolfgang Lederer, Daniel Heidi, Tristan Anselm Kuder, Daniel Paech, David Bonekamp, Alexander Radbruch, Stefan Delorme, Heinz-Peter Schlemmer, Franziska Steudle, and Klaus H. Maier-Hein. **Revealing Hidden Potentials of the Q-Space Signal in Breast Cancer**. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Lecture Notes in Computer Science, pages

664–671, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66182-7. doi: 10.1007/978-3-319-66182-7_76.

Sarthak Jain and Byron C. Wallace. **Attention is not Explanation**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL `https://aclanthology.org/N19-1357`.

Manuela Jeyaraj and Dharshana Kasthurirathna. **MNet-Sim: A Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity**. *International Journal of Engineering Trends and Technology*, 69:181–189, 07 2021. doi: 10.14445/22315381/IJETT-V69I7P225.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. **SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.197. URL `https://aclanthology.org/2020.acl-main.197`.

Meirui Jiang, Zirui Wang, and Qi Dou. **HarmoFL: Harmonizing Local and Global Drifts in Federated Learning on Heterogeneous Medical Images**. *AAAI Conference on Artificial Intelligence*, 2022.

Min Jiang, Yonghui Wu, Anushi Shah, Priyanka Priyanka, Joshua C Denny, and Hua Xu. **Extracting and standardizing medication information in clinical text - the MedEx-UIMA system**. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2014:3742, 2014. ISSN 2153-4063. URL `https://europepmc.org/articles/PMC4419757`.

Joint Imaging Platform. **Joint Imaging Platform**. https://jip.dktk.dkfz.de/jiphomepage/, 2023. Accessed: January 20, 2023.

Klaus Kades, Jan Sellner, Gregor Koehler, Peter M Full, T Y Emmy Lai, Jens Kleesiek, and Klaus H Maier-Hein. **Adapting Bidirectional Encoder Representations from Transformers (BERT) to Assess Clinical Semantic Textual Similarity: Algorithm Development and Validation Study**. *JMIR Med Inform*, 9(2):e22795, Feb 2021. ISSN 2291-9694. doi: 10.2196/22795. URL `https://medinform.jmir.org/2021/2/e22795`.

Klaus Kades, Jonas Scherer, Jan Scholtyssek, Tobias Penzkofer, Marco Nolden, and Klaus Maier-Hein. **Efficient DICOM Image Tagging and Cohort Curation Within Kaapana**. In Klaus Maier-Hein, Thomas M. Deserno, Heinz Handels, Andreas Maier, Christoph Palm, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2022*, pages 279–284, Wiesbaden, 2022a. Springer Fachmedien Wiesbaden. ISBN 978-3-658-36932-3.

Klaus Kades, Jonas Scherer, Maximilian Zenk, Marius Kempf, and Klaus Maier-Hein. **Towards Real-World Federated Learning inăMedical Image Analysis Using Kaapana**. In Shadi Albarqouni, Spyridon Bakas, Sophia Bano, M. Jorge Cardoso, Bishesh Khanal, Bennett Landman, Xiaoxiao Li, Chen Qin, Islem Rekik, Nicola Rieke, Holger Roth, Debdoot Sheet, and Daguang Xu, editors, *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health*, pages 130–140, Cham, 2022b. Springer Nature Switzerland. ISBN 978-3-031-18523-6.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawit, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. DOliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021.

Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. **Secure, privacy-preserving and federated machine learning in medical imaging**. *Nature Machine Intelligence*, 2(6):305–311, Jun 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0186-1. URL `https://doi.org/10.1038/s42256-020-0186-1`.

Ning Kang, Bharat Singh, Zubair Afzal, Erik M. van Mulligen, and Jan Kors. **Using rule-based natural language processing to improve disease normalization in biomedical text**. *Journal of the American Medical Informatics Association : JAMIA*, 20, 10 2012. doi: 10.1136/amiajnl-2012-001173.

Benjamin H. Kann, Ahmed Hosny, and Hugo J.W.L. Aerts. **Artificial intelligence for clinical oncology**. *Cancer Cell*, 39(7):916–927, 2021. ISSN 1535-6108. doi:

https://doi.org/10.1016/j.ccell.2021.04.002. URL `https://www.sciencedirect.com/science/article/pii/S1535610821002105`.

Kenneth L. Kehl, Haitham Elmarakeby, Mizuki Nishino, Eliezer M. Van Allen, Eva M. Lepisto, Michael J. Hassett, Bruce E. Johnson, and Deborah Schrag. **Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports**. 5(10):1421–1429, 2019. ISSN 2374-2445. doi: 10.1001/jamaoncol.2019.1800.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. **CTRL - A Conditional Transformer Language Model for Controllable Generation**. *arXiv preprint arXiv:1909.05858*, 2019.

Keycloak. **Keycloak**. https://www.keycloak.org/, 2023. Accessed: January 18, 2023.

Jens Kleesiek, Jacob M. Murray, Christian Strack, Sebastian Prinz, Georgios Kaissis, and Rickmer Braren. **Künstliche Intelligenz und maschinelles Lernen in der onkologischen Bildgebung**. *Der Pathologe*, 41(6):649–658, Nov 2020. ISSN 1432-1963. doi: 10.1007/s00292-020-00827-3. URL `https://doi.org/10.1007/s00292-020-00827-3`.

Matthias Klumpp, Marcus Hintze, Milla Immonen, Francisco Ródenas-Rigla, Francesco Pilati, Fernando Aparicio-Martínez, Dilay Çelebi, Thomas Liebig, Mats Jirstrand, Oliver Urbann, Marja Hedman, Jukka A. Lipponen, Silvio Bicciato, Anda-Petronela Radan, Bernardo Valdivieso, Wolfgang Thronicke, Dimitrios Gunopulos, and Ricard Delgado-Gonzalo. **Artificial Intelligence for Hospital Health Care: Application Cases and Answers to Challenges in European Hospitals**. *Healthcare*, 9(8), 2021. ISSN 2227-9032. doi: 10.3390/healthcare9080961. URL `https://www.mdpi.com/2227-9032/9/8/961`.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. **Attention is Not Only a Weight: Analyzing Transformers with Vector Norms**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.574. URL `https://aclanthology.org/2020.emnlp-main.574`.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. **Automatic ICD-10 classification of cancers from free-text death certificates**. *International Journal of Medical Informatics*, 84(11):956–965, 2015. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2015.08.004. URL `https://www.sciencedirect.com/science/article/pii/S1386505615300289`.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. **Revealing the Dark Secrets of BERT**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL `https://aclanthology.org/D19-1445`.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. **Improving Abstraction in Text Summarization**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1207. URL `https://aclanthology.org/D18-1207`.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. **Evaluating the Factual Consistency of Abstractive Text Summarization**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL `https://aclanthology.org/2020.emnlp-main.750`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**. *arXiv e-prints*, art. arXiv:1909.11942, September 2019.

Andras Lasso. **AI-assisted Segmentation Extension - Announcements - 3D Slicer Community**. https://discourse.slicer.org/t/ai-assisted-segmentation-extension/9536, December 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*, 36(4):1234–1240, 2020.

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. **Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks**. *arXiv e-prints*, art. arXiv:1511.06314, November 2015.

Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C. Vilanova, Paul M. Walker, and Fabrice Meriaudeau. **Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review**. *Computers in Biology and Medicine*, 60:8–31, 2015. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2015.02.009. URL `https://www.sciencedirect.com/science/article/pii/S001048251500058X`.

Junyi Li, Xuejie Zhang, and Xiaobing Zhou. **ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study**. *JMIR Med Inform*, 9(1):e23086, Jan 2021a. ISSN 2291-9694. doi: 10.2196/23086. URL `http://medinform.jmir.org/2021/1/e23086/`.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. **A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection**. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021b. doi: 10.1109/TKDE.2021.3124599.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. **Federated Learning: Challenges, Methods, and Future Directions**. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. doi: 10.1109/MSP.2020.2975749.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. **Improving Neural Abstractive Document Summarization with Explicit Information Selection Modeling**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1205. URL `https://aclanthology.org/D18-1205`.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. **Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation**. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL `https://proceedings.neurips.cc/paper/2018/file/e07413354875be01a996dc560274708e-Paper.pdf`.

Siting Liang. **Summarizing German Radiology Findings for Cancer Patients**, 2021. Master thesis.

Siting Liang, Klaus Kades, Matthias Fink, Peter Full, Tim Weber, Jens Kleesiek, Michael Strube, and Klaus Maier-Hein. **Fine-tuning BERT Models for Summarizing German Radiology Findings**. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 30–40, Seattle, WA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.4. URL `https://aclanthology.org/2022.clinicalnlp-1.4`.

Chin-Yew Lin. **ROUGE: A Package for Automatic Evaluation of Summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

Nathaniel Linna and Charles E. Kahn. **Applications of natural language processing in radiology: A systematic review**. *International Journal of Medical Informatics*, 163:104779, 2022. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2022.104779. URL `https://www.sciencedirect.com/science/article/pii/S1386505622000934`.

Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. **Large Scale Diagnostic Code Classification for Medical Patient Records**. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008. URL `https://aclanthology.org/I08-2125`.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. **Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge**. *Medical Image Analysis*, 18(2):359–373, 2014. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2013.12.002. URL `https://www.sciencedirect.com/science/article/pii/S1361841513001734`.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. **Clinically Accurate Chest X-Ray Report Generation**. *arXiv e-prints*, art. arXiv:1904.02633, April 2019.

Quande Liu, Qi Dou, and Pheng-Ann Heng. **Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains**. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 475–485, Cham, 2020a. Springer International Publishing. ISBN 978-3-030-59713-9.

Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. **Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data**. *IEEE Transactions on Medical Imaging*, 2020b.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. **FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space**. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Sijia Liu, Yanshan Wang, and Hongfang Liu. **Selected articles from the BioCreative/OHNLP challenge 2018**. 19(10):262, 2019a. ISSN 1472-6947. doi: 10.1186/s12911-019-0994-6. URL `https://doi.org/10.1186/s12911-019-0994-6`.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. **Multi-Task Deep Neural Networks for Natural Language Understanding**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics, 2019b. doi: 10.18653/v1/P19-1441. URL `https://www.aclweb.org/anthology/P19-1441`.

Yang Liu and Mirella Lapata. **Text Summarization with Pretrained Encoders**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL `https://aclanthology.org/D19-1387`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. *ArXiv*, abs/1907.11692, 2019c.

Justin Lovelace and Bobak Mortazavi. **Learning to Generate Clinically Coherent Chest X-Ray Reports**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.110. URL `https://aclanthology.org/2020.findings-emnlp.110`.

Ludovic Fernandez. **Traefik**. https://doc.traefik.io/traefik/, 2023. Accessed: January 18, 2023.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. **Ontology-Aware Clinical Abstractive Summarization**. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 10131016, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331319. URL `https://doi.org/10.1145/3331184.3331319`.

Kirti Magudia, Christopher P. Bridge, Katherine P. Andriole, and Michael H. Rosenthal. **The Trials and Tribulations of Assembling Large Medical Imaging Datasets for Machine Learning Applications**. *Journal of Digital Imaging*, 34(6):1424–1429, Dec 2021. ISSN 1618-727X. doi: 10.1007/s10278-021-00505-7. URL `https://doi.org/10.1007/s10278-021-00505-7`.

Diwakar Mahajan, Ananya Poddar, Jennifer J Liang, Yen-Ting Lin, John M Prager, Parthasarathy Suryanarayanan, Preethi Raghavan, and Ching-Huei Tsou. **Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning**. *JMIR Med Inform*, 8 (11):e22508, Nov 2020. ISSN 2291-9694. doi: 10.2196/22508. URL `http://medinform.jmir.org/2020/11/e22508/`.

Goutam Majumder, Dr. Partha Pakray, Alexander Gelbukh, and David Pinto. **Semantic Textual Similarity Methods, Tools, and Applications: A Survey**. 20:647–665, 2016. doi: 10.13053/CyS-20-4-2506.

Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. **SMART on FHIR: a standards-based, interoperable apps platform for electronic health records**. *Journal of the American Medical Informatics Association*, 23(5):899–908, 02 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocv189. URL `https://doi.org/10.1093/jamia/ocv189`.

Leland McInnes, John Healy, and James Melville. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. *arXiv e-prints*, art. arXiv:1802.03426, February 2018.

Medical Segmentation Decathlon. **Medical Segmentation Decathlon**. http://medicaldecathlon.com/results/, 2023. Accessed: January 20, 2023.

Peter Mell and Tim Grance. **The NIST Definition of Cloud Computing**. Technical Report NIST Special Publication (SP) 800-145, National Institute of Standards and Technology, September 2011.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çaatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. **The Multimodal Brain**

Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Dirk Merkel. **Docker: Lightweight Linux Containers for Consistent Development and Deployment**. *Linux journal*, 2014(239):2, 2014.

Dieter Meschede. **Gerthsen Physik**. Springer Spektrum Berlin, Heidelberg, 25 edition, 2015. ISBN 978-3-662-45976-8. doi: https://doi.org/10.1007/978-3-662-45977-5.

Microk8s. **Microk8s**. https://microk8s.io/, 2023. Accessed: January 20, 2023.

Microsoft. **Windows Server**. Microsoft, 2023. Accessed: January 20, 2023.

Microsoft. **Cloud Computing Services | Microsoft Azure**. https://azure.microsoft.com/en-us/, 2023. Accessed: January 18, 2023.

Rada Mihalcea and Paul Tarau. **TextRank: Bringing Order into Text**. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-3252`.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. **Efficient Estimation of Word Representations in Vector Space**. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

Derek Miller. **Leveraging BERT for Extractive Text Summarization on Lectures**. *arXiv e-prints*, art. arXiv:1906.04165, June 2019.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. **Deep Learning–Based Text Classification: A Comprehensive Review**. *ACM Comput. Surv.*, 54(3), apr 2021. ISSN 0360-0300. doi: 10.1145/3439726. URL `https://doi.org/10.1145/3439726`.

MinIO. **MinIO**. https://min.io, 2023. Accessed: January 18, 2023.

Mint Medical. **Mint Medical**. https://mint-medical.com/, 2023. Accessed: January 20, 2023.

Andriy Mulyar, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. **Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models**. *arXiv e-prints*, art. arXiv:1910.13664, October 2019.

Bloch N, Madabhushi A, Huisman H, Freymann J, Kirby J, Grauer M, Enquobahrie A, Jaffe C, Clarke L, and Farahani K. **NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. The Cancer Imaging Archive.**, 2015.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. **doccano: Text Annotation Tool for Human**, 2018. URL `https://github.com/doccano/doccano`. Software available from https://github.com/doccano/doccano.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. **Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond**. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL `https://aclanthology.org/K16-1028`.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. **Classify or Select: Neural Architectures for Extractive Document Summarization**. *arXiv e-prints*, art. arXiv:1611.04244, November 2016.

National NLP Clinical Challenges. **National NLP Clinical Challenges**. https://n2c2.dbmi.hms.harvard.edu/2019-challenge, 2023. Accessed: January 20, 2023.

NEMA. **DICOM - DIMSE-C**. http://dicom.nema.org/dicom/2013/output/chtml/-part07/sect_9.3.html, 2023. Accessed: January 18, 2023.

Medical Imaging & Technology Alliance NEMA. **NEMA PS3 / ISO 12052, Digital Imaging and Communications in Medicine (DICOM) Standard**, 2021.

Elisa Nguyen, Daphne Theodorakopoulos, Shreyasi Pathak, Jeroen Geerdink, Onno Vijlbrief, Maurice van Keulen, and Christin Seifert. **A Hybrid Text Classification and Language Generation Model for Automated Summarization of Dutch Breast Cancer Radiology Reports**. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, pages 72–81, 2020. doi: 10.1109/CogMI50398.2020.00019.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. **Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark**. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.3. URL `https://aclanthology.org/2021.nllp-1.3`.

nnU-Net. **nnU-Net**. https://github.com/MIC-DKFZ/nnUNet, 2023. Accessed: January 20, 2023.

J. Martijn Nobel, Ellen M. Kok, and Simon G. F. Robben. **Redefining the structure of structured reporting in radiology**. *Insights into Imaging*, 11(1):10, Feb 2020.

ISSN 1869-4101. doi: 10.1186/s13244-019-0831-6. URL `https://doi.org/10.1186/s13244-019-0831-6`.

Marco Nolden, Sascha Zelzer, Alexander Seitel, Diana Wald, Michael Müller, Alfred M. Franz, Daniel Maleike, Markus Fangerau, Matthias Baumhauer, Lena Maier-Hein, Klaus H. Maier-Hein, Hans-Peter Meinzer, and Ivo Wolf. **The Medical Imaging Interaction Toolkit: challenges and advances : 10 years of open-source development**. *Int J Comput Assist Radiol Surg*, 8(4):607–620, July 2013. ISSN 1861-6429. doi: 10.1007/s11548-013-0840-8.

NVIDIA Clara. **NVIDIA Clara**. https://docs.nvidia.com/clara/, 2023. Accessed: January 20, 2023.

Michael Oakes. **Statistics for Corpus Linguistics**. Edinburgh University Press, Edinburgh, 1998. ISBN 9781474471381. doi: doi:10.1515/9781474471381. URL `https://doi.org/10.1515/9781474471381`.

OAuth2 Proxy. **OAuth2 Proxy**. https://oauth2-proxy.github.io/oauth2-proxy/, 2023. Accessed: January 18, 2023.

OpenMined. **OpenMined**. https://github.com/OpenMined, 2023. Accessed: January 20, 2023.

David Opitz and Richard Maclin. **Popular Ensemble Methods: An Empirical Study**. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

Optuna. **Optuna**. https://github.com/optuna/optuna, 2023. Accessed: January 20, 2023.

PaddlePaddle. **PaddleFL**. https://github.com/PaddlePaddle/PaddleFL, 2023. Accessed: January 20, 2023.

pattern. **pattern**. https://github.com/clips/pattern, 2023. Accessed: January 20, 2023.

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*, 12:2825–2830, 2011.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. **Scikit-learn: Machine Learning in Python**. 12(85):2825–2830, 2018. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

Nathan Peiffer-Smadja, Redwan Maatoug, François-Xavier Lescure, Eric D'Ortenzio, Joëlle Pineau, and Jean-Rémi King. **Machine Learning for COVID-19 needs global collaboration and data-sharing**. *Nature Machine Intelligence*, 2(6):293–294, Jun 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0181-6. URL `https://doi.org/10.1038/s42256-020-0181-6`.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. **GloVe: Global Vectors for Word Representation**. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Podman community. **Podman**. https://podman.io/, 2023. Accessed: January 18, 2023.

Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. **Natural Language Processing in Radiology: A Systematic Review**. *Radiology*, 279(2): 329–343, 2016. doi: 10.1148/radiol.16142770. URL `https://doi.org/10.1148/radiol.16142770`. PMID: 27089187.

PostgreSQL. **PostgreSQL**. https://www.postgresql.org/, 2023. Accessed: January 20, 2023.

Prayitno, Chi-Ren Shyu, Karisma Trinanda Putra, Hsing-Chung Chen, Yuan-Yu Tsai, K. S. M. Tozammel Hossain, Wei Jiang, and Zon-Yin Shae. **A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications**. *Applied Sciences*, 11(23), 2021. ISSN 2076-3417. doi: 10.3390/app112311191. URL `https://www.mdpi.com/2076-3417/11/23/11191`.

Prometheus. **Alertmanager**. https://github.com/prometheus/alertmanager, 2023a. Accessed: January 20, 2023.

Prometheus. **Prometheus**. https://prometheus.io/, 2023b. Accessed: January 20, 2023.

PROMISE12. **PROMISE12**. https://promise12.grand-challenge.org/, 2023. Accessed: January 20, 2023.

PyContractions. **PyContractions**. https://github.com/ian-beaver/pycontractions, 2023. Accessed: January 20, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. **Language Models are Unsupervised Multitask Learners**. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. **Exploring the Limits of Transfer**

Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. **AI in health and medicine**. *Nature Medicine*, 28(1):31–38, Jan 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01614-0. URL `https://doi.org/10.1038/s41591-021-01614-0`.

RECIST. **RECIST**. https://recist.eortc.org/, 2023. Accessed: January 20, 2023.

David K. Rensin. **Kubernetes - Scheduling the Future at Cloud Scale**. In *OSCON 2015*, page All. O'Reilly Media, Inc., 1005 Gravenstein Highway North Sebastopol, CA 95472, 2015.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. **On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities**. *Radiology: Artificial Intelligence*, 2(3):e190043, 2020. doi: 10.1148/ryai.2020190043. URL `https://doi.org/10.1148/ryai.2020190043`. PMID: 32510054.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. **The future of digital health with federated learning**. *npj Digital Medicine*, 3(1):119, Sep 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00323-1. URL `https://doi.org/10.1038/s41746-020-00323-1`.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Holger R. Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C. Bizzo, Yuhong Wen, Varun Buch, Meesam Shah, Felipe Kitamura, Matheus Mendonça, Vitor Lavor, Ahmed Harouni, Colin Compas, Jesse Tetreault, Prerna Dogra, Yan Cheng, Selnur Erdal, Richard White, Behrooz Hashemian, Thomas Schultz, Miao Zhang, Adam McCarthy, B. Min Yun, Elshaimaa Sharaf, Katharina V. Hoebel, Jay B. Patel, Bryan Chen, Sean Ko, Evan Leibovitz, Etta D. Pisano, Laura Coombs, Daguang Xu, Keith J. Dreyer, Ittai Dayan, Ram C. Naidu, Mona Flores, Daniel Rubin, and Jayashree Kalpathy-Cramer. **Federated Learning for Breast Density Classification: A Real-World**

Implementation. In Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu, editors, *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 181–191, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60548-3.

Holger R. Roth, Dong Yang, Wenqi Li, Andriy Myronenko, Wentao Zhu, Ziyue Xu, Xiaosong Wang, and Daguang Xu. **Federated Whole Prostate Segmentation in MRI with Personalized Neural Architectures**. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 357–366, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. **Leveraging Pre-trained Checkpoints for Sequence Generation Tasks**. *Transactions of the Association for Computational Linguistics*, 8:264–280, 2020. doi: 10.1162/tacl_a_00313. URL `https://aclanthology.org/2020.tacl-1.18`.

Daniel L. Rubin, Cesar Rodriguez, Priyanka Shah, and Chris Beaulieu. **iPad: Semantic Annotation and Markup of Radiological Images**. *AMIA Annual Symposium Proceedings*, 2008:626–630, 2008. ISSN 1942-597X. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655990/`.

Daniel L. Rubin, Debra Willrett, Martin J. O’Connor, Cleber Hage, Camille Kurtz, and Dilvan A. Moreira. **Automated Tracking of Quantitative Assessments of Tumor Burden in Clinical Trials**. *Translational Oncology*, 7(1):23–35, 2014. ISSN 1936-5233. doi: https://doi.org/10.1593/tlo.13796. URL `https://www.sciencedirect.com/science/article/pii/S1936523314800041`. The Quantitative Imaging Network.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. **Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses**. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. **ImageNet Large Scale Visual Recognition Challenge**. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL `https://doi.org/10.1007/s11263-015-0816-y`.

Karthik V Sarma, Stephanie Harmon, Thomas Sanford, Holger R Roth, Ziyue Xu, Jesse Tetreault, Daguang Xu, Mona G Flores, Alex G Raman, Rushikesh Kulkarni, Bradford J Wood, Peter L Choyke, Alan M Priester, Leonard S Marks, Steven S Raman, Dieter Enzmann, Baris Turkbey, William Speier, and Corey W Arnold. **Federated learning improves site performance in multicenter deep learning without data sharing**. *Journal of the American Medical Informatics Association*, 28(6): 1259–1264, 02 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa341. URL `https://doi.org/10.1093/jamia/ocaa341`.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. **One Million Posts: A Data Set of German Online Discussions**. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan, August 2017. doi: 10.1145/3077136.3080711.

Jonas Scherer. **Decentralized Infrastructure for Medical Image Analysis**, 2022. PhD thesis.

Jonas Scherer, Marco Nolden, Jens Kleesiek, Jasmin Metzger, Klaus Kades, Verena Schneider, Michael Bach, Oliver Sedlaczek, Andreas M. Bucher, Thomas J. Vogl, Frank Grünwald, Jens-Peter Kühn, Ralf-Thorsten Hoffmann, Jörg Kotzerke, Oliver Bethge, Lars Schimmöller, Gerald Antoch, Hans-Wilhelm Müller, Andreas Daul, Konstantin Nikolaou, Christian la Fougère, Wolfgang G. Kunz, Michael Ingrisch, Balthasar Schachtner, Jens Ricke, Peter Bartenstein, Felix Nensa, Alexander Radbruch, Lale Umutlu, Michael Forsting, Robert Seifert, Ken Herrmann, Philipp Mayer, Hans-Ulrich Kauczor, Tobias Penzkofer, Bernd Hamm, Winfried Brenner, Roman Kloeckner, Christoph Düber, Mathias Schreckenberger, Rickmer Braren, Georgios Kaissis, Marcus Makowski, Matthias Eiber, Andrei Gafita, Rupert Trager, Wolfgang A. Weber, Jakob Neubauer, Marco Reisert, Michael Bock, Fabian Bamberg, Jürgen Hennig, Philipp Tobias Meyer, Juri Ruf, Uwe Haberkorn, Stefan O. Schoenberg, Tristan Kuder, Peter Neher, Ralf Floca, Heinz-Peter Schlemmer, and Klaus Maier-Hein. **Joint Imaging Platform for Federated Clinical Data Analytics**. *JCO Clinical Cancer Informatics*, 4:1027–1038, November 2020. doi: 10.1200/CCI.20.00045.

Jonas Scherer, Klaus Kades, Hanno Gao, Ralf Floca, Peter Neher, Marco Nolden, and Klaus Maier-Hein. **Kaapana**. https://github.com/kaapana/kaapana, 2023. Accessed: January 20, 2023.

Sarah Schuhegger. **Body Part Regression for CT Images**. *arXiv e-prints*, art. arXiv:2110.09148, October 2021.

scikit-learn. **CalibratedClassifierCV**. https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html, 2023. Accessed: January 20, 2023.

Abigail See, Peter J. Liu, and Christopher D. Manning. **Get To The Point: Summarization with Pointer-Generator Networks**. *arXiv e-prints*, art. arXiv:1704.04368, April 2017.

segtok. **segtok**. https://github.com/fnl/segtok, 2023. Accessed: January 20, 2023.

Sofia Serrano and Noah A. Smith. **Is Attention Interpretable?** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL `https://aclanthology.org/P19-1282`.

Farah Shamout, Tingting Zhu, and David A. Clifton. **Machine Learning for Clinical Outcome Prediction**. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2021. doi: 10.1109/RBME.2020.3007816.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Colen, and Spyridon Bakas. **Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data**. *Scientific Reports*, 10(1):12598, Jul 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69250-1. URL `https://doi.org/10.1038/s41598-020-69250-1`.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. **Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis**. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018. doi: 10.1109/JBHI.2017.2767063.

Manjil Shrestha. **Development of a Language Model for Medical Domain**. Masterthesis, Hochschule Rhein-Waal, 2021.

SMART. **2019 SMART Flat FHIR/Bulk Data Meeting**. https://smarthealthit.org/2019-smart-flat-fhir-bulk-data-meeting/, 2019.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. **Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL `https://aclanthology.org/2020.emnlp-main.117`.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. **Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.172. URL `https://aclanthology.org/2020.acl-main.172`.

KAREN SPARCK JONES. **A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL**. *Journal of Documentation*, 28(1):11–21, Jan 1972. ISSN 0022-0418. doi: 10.1108/eb026526. URL `https://doi.org/10.1108/eb026526`.

Tobias Stein, Jasmin Metzger, Jonas Scherer, Fabian Isensee, Tobias Norajitra, Jens Kleesiek, Klaus Maier-Hein, and Marco Nolden. **Efficient Web-Based Review for Automatic Segmentation of Volumetric DICOM Images**. In Heinz Handels, Thomas M. Deserno, Andreas Maier, Klaus Hermann Maier-Hein, Christoph Palm, and Thomas Tolxdorff, editors, *Bildverarbeitung für die Medizin 2019*, pages 158–163, Wiesbaden, 2019. Springer Fachmedien Wiesbaden. ISBN 978-3-658-25326-4.

Jackson M. Steinkamp, Charles M. Chambers, Darco Lalevic, Hanna M. Zafar, and Tessa S. Cook. **Automated Organ-Level Classification of Free-Text Pathology Reports to Support a Radiology Follow-up Tracking Engine**. *Radiology: Artificial Intelligence*, 1(5):e180052, 2019. doi: 10.1148/ryai.2019180052. URL `https://doi.org/10.1148/ryai.2019180052`. PMID: 33937800.

STSbenchmark. **STSbenchmark**. https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark, 2023. Accessed: January 20, 2023.

Jimeng Sun and Chandan K. Reddy. **Big Data Analytics for Healthcare**. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, page 1525, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2506178. URL `https://doi.org/10.1145/2487575.2506178`.

Xiaobing Sun and Wei Lu. **Understanding Attention for Text Classification**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.312. URL `https://aclanthology.org/2020.acl-main.312`.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. **ERNIE 2.0: A Continual Pre-training Framework for Language Understanding**. 2019. URL `http://arxiv.org/abs/1907.12412`. version: 1.

TCIA. **NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures (ISBI-MR-Prostate-2013)**. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21267207, 2023. Accessed: January 20, 2023.

teem. **Nrrd**. http://teem.sourceforge.net/nrrd/index.html, 2023. Accessed: January 18, 2023.

TensorFlow. **TensorFlow Federated**. https://www.tensorflow.org/federated, 2023. Accessed: January 20, 2023.

The Linux Foundation. **Helm**. https://helm.sh/, 2023. Accessed: January 18, 2023.

The radiology cooperation in NUM. **RACOON**. https://racoon.network/, 2023. Accessed: January 20, 2023.

Daniel Shu Wei Ting, Lawrence Carin, Victor Dzau, and Tien Y. Wong. **Digital technology and COVID-19**. *Nature Medicine*, 26(4):459–461, Apr 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0824-5. URL `https://doi.org/10.1038/s41591-020-0824-5`.

Daiju Ueda, Akitoshi Shimazaki, and Yukio Miki. **Technical and clinical overview of deep learning in radiology**. *Japanese Journal of Radiology*, 37(1):15–33, Jan 2019. ISSN 1867-108X. doi: 10.1007/s11604-018-0795-3. URL `https://doi.org/10.1007/s11604-018-0795-3`.

UK biobank. **UK biobank**. https://www.ukbiobank.ac.uk/, 2023. Accessed: January 20, 2023.

Trinity Urban, Erik Ziegler, Rob Lewis, Chris Hafey, Cheryl Sadow, Annick D. Van den Abbeele, and Gordon J. Harris. **LesionTracker: Extensible Open-Source Zero-Footprint Web Viewer for Cancer Imaging Research and Clinical Trials**. *Cancer Research*, 77(21):e119–e122, 10 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0334. URL `https://doi.org/10.1158/0008-5472.CAN-17-0334`.

Laurens van der Maaten and Geoffrey Hinton. **Visualizing Data using t-SNE**. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

Shikhar Vashishth, Manik Bhandari, Prateek Yadav, Piyush Rai, Chiranjib Bhattacharyya, and Partha Talukdar. **Incorporating Syntactic and Semantic Information in Word Embeddings using Graph Convolutional Networks**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3308–3318, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1320. URL `https://aclanthology.org/P19-1320`.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. **Attention Interpretability Across NLP Tasks**. *arXiv e-prints*, art. arXiv:1909.11218, September 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. **Attention Is All You Need**. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

S. Velupillai, D. Mowery, B. R. South, M. Kvist, and H. Dalianis. **Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis**. *Yearb Med Inform*, 24(01):183–193, 10.03.2018. 183.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. **Pointer Networks**. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/29921001f2f04bd3baee84a12e98098f-Paper.pdf`.

Vue.js. **Vue.js**. https://vuejs.org/, 2023. Accessed: January 20, 2023.

Vuetify . **Vuetify**. https://vuetifyjs.com/, 2023. Accessed: January 20, 2023.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. **StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding**. *arXiv e-prints*, art. arXiv:1908.04577, August 2019.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. **MedSTS: a resource for clinical semantic textual similarity**. 2018a. doi: 10.1007/s10579-018-9431-1.

Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. **Overview of BioCreative/OHNLP Challenge 2018 Task 2: Clinical Semantic Textual Similarity**. 2018b. doi: 10.13140/RG.2.2.26682.24006.

Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. **The 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity: Overview**. *JMIR Med Inform*, 8(11):e23375, Nov 2020. ISSN 2291-9694. doi: 10.2196/23375. URL `http://medinform.jmir.org/2020/11/e23375/`.

David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. **Clinical applications of machine learning algorithms: beyond the black box**. *BMJ*, 364, 2019. ISSN 0959-8138. doi: 10.1136/bmj.l886. URL `https://www.bmj.com/content/364/bmj.l886`.

TF Weber, M Spurny, FC Hasse, O Sedlaczek, GM Haag, C Springfeld, T Mokry, D Jäger, HU Kauczor, and AK Berger. **Improving radiologic communication in oncology:**

a single-centre experience with structured reporting for cancer patients. *Insights Imaging*, 11(1):106, Sep 29 2020. doi: 10.1186/s13244-020-00907-1.

Sarah Wiegreffe and Yuval Pinter. **Attention is not not Explanation**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL `https://aclanthology.org/D19-1002`.

Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. **Methods and open-source toolkit for analyzing and visualizing challenge results**. *Scientific Reports*, 11(1):2369, Jan 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82017-6. URL `https://doi.org/10.1038/s41598-021-82017-6`.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data*, 3(1):160018, Mar 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL `https://doi.org/10.1038/sdata.2016.18`.

Martin J. Willemink, Wojciech A. Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R. Folio, Ronald M. Summers, Daniel L. Rubin, and Matthew P. Lungren. **Preparing Medical Imaging Data for Machine Learning**. *Radiology*, 295(1):4–15, 2020. doi: 10.1148/radiol.2020192224. URL `https://doi.org/10.1148/radiol.2020192224`. PMID: 32068507.

Ivo Wolf, Marcus Vetter, Ingmar Wegner, Marco Nolden, Thomas Bottger, Mark Hastenteufel, Max Schobinger, Tobias Kunert, and Hans-Peter Meinzer. **The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK**. In Robert L. Galloway Jr., editor, *Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display*, volume 5367, pages 16 – 27. International Society for Optics and Photonics, SPIE, 2004. doi: 10.1117/12.535112. URL `https://doi.org/10.1117/12.535112`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. **Transformers: State-of-the-Art Natural Language Processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. **Hi-Transformer: Hierarchical Interactive Transformer for Efficient and Effective Long Document Modeling**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 848–853, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.107. URL `https://aclanthology.org/2021.acl-short.107`.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. **MedEx: a medication information extraction system for clinical narratives**. *Journal of the American Medical Informatics Association*, 17 (1):19–24, 01 2010. ISSN 1067-5027. doi: 10.1197/jamia.M3378. URL `https://doi.org/10.1197/jamia.M3378`.

Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. **Federated Learning for Healthcare Informatics**. *Journal of Healthcare Informatics Research*, 5(1):1–19, Mar 2021. ISSN 2509-498X. doi: 10.1007/s41666-020-00082-4. URL `https://doi.org/10.1007/s41666-020-00082-4`.

Dong Yang, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R. Roth, Stephanie Harmon, Sheng Xu, Baris Turkbey, Evrim Turkbey, Xiaosong Wang, Wentao Zhu, Gianpaolo Carrafiello, Francesca Patella, Maurizio Cariati, Hirofumi Obinata, Hitoshi Mori, Kaku Tamura, Peng An, Bradford J. Wood, and Daguang Xu. **Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan**. *Medical Image Analysis*, 70:101992, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.101992. URL `https://www.sciencedirect.com/science/article/pii/S1361841521000384`.

Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. **Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models**. *JMIR Med Inform*, 8(11):e19735, Nov 2020. ISSN 2291-9694. doi: 10.2196/19735. URL `http://medinform.jmir.org/2020/11/e19735/`.

Wen-wai Yim, Meliha Yetisgen, William P. Harris, and Sharon W. Kwan. **Natural Language Processing in Oncology: A Review**. *JAMA Oncology*, 2(6):797–804, 06 2016. ISSN 2374-2437. doi: 10.1001/jamaoncol.2016.0213. URL `https://doi.org/10.1001/jamaoncol.2016.0213`.

Ying Yu, Min Li, Liangliang Liu, Yaohang Li, and Jianxin Wang. **Clinical big data and deep learning: Applications, challenges, and future outlooks**. *Big Data Mining and Analytics*, 2(4):288–305, 2019. doi: 10.26599/BDMA.2019.9020007.

G. UDNY YULE. **ON SENTENCE- LENGTH AS A STATISTICAL CHARACTERISTIC OF STYLE IN PROSE: WITH APPLICATION TO TWO CASES OF DISPUTED AUTHORSHIP**. *Biometrika*, 30(3-4):363–390, 01 1939. ISSN 0006-3444. doi: 10.1093/biomet/30.3-4.363. URL `https://doi.org/10.1093/biomet/30.3-4.363`.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. **Big bird: Transformers for longer sequences**. *Advances in Neural Information Processing Systems*, 33, 2020.

John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar, and Eric Karl Oermann. **Natural Languagebased Machine Learning Models for the Annotation of Clinical Radiology Reports**. *Radiology*, 287(2):570–580, 2018. doi: 10.1148/radiol.2018171093. URL `https://doi.org/10.1148/radiol.2018171093`. PMID: 29381109.

Rui Zhang, Serguei V. Pakhomov, Janet T. Lee, and Genevieve B. Melton. **Using language models to identify relevant new information in inpatient clinical notes**. 2014:1268–1276, 2014. ISSN 1942-597X.

Shuang Zhang, Xuefeng Zheng, and Changjun Hu. **A survey of semantic similarity and its application to social network analysis**. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2362–2367, 2015. doi: 10.1109/BigData.2015.7364028.

Xingxing Zhang, Furu Wei, and Ming Zhou. **HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization**. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5059–5069, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1499. URL `https://aclanthology.org/P19-1499`.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. **Learning to Summarize Radiology Findings**. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5623. URL `https://aclanthology.org/W18-5623`.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. **Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.458. URL `https://aclanthology.org/2020.acl-main.458`.

Xiangxin Zhu, Carl Vondrick, Charless C. Fowlkes, and Deva Ramanan. **Do We Need More Training Data?** *International Journal of Computer Vision*, 119(1): 76–92, Aug 2016. ISSN 1573-1405. doi: 10.1007/s11263-015-0812-2. URL `https://doi.org/10.1007/s11263-015-0812-2`.

Erik Ziegler, Trinity Urban, Danny Brown, James Petts, Steve D. Pieper, Rob Lewis, Chris Hafey, and Gordon J. Harris. **Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research**. *JCO Clinical Cancer Informatics*, (4): 336–345, 2020. doi: 10.1200/CCI.19.00131. URL `https://doi.org/10.1200/CCI.19.00131`. PMID: 32324447.

Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. **Medical imaging deep learning with differential privacy**. *Scientific Reports*, 11(1):13524, Jun 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-93030-0. URL `https://doi.org/10.1038/s41598-021-93030-0`.

Current Challenges in the Application of Algorithms in Multi-institutional Clinical Settings

Ph. D. Thesis

Supervised by Prof. Dr. Klaus H. Maier-Hein

This work has been set using LaTeX

Color Scheme: `https://personal.sron.nl/~pault/`