



Technische Universität München
TUM School of Computation, Information and Technology

Multimodal Deep Learning for Holistic Clinical Decision and Reasoning Support

Matthias Fabian Keicher

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Rückert

Prüfende der Dissertation:

1. Prof. Dr. Nassir Navab
2. Prof. Dr.-Ing. Andreas Maier
3. Prof. Pranav Rajpurka, Ph.D.

Die Dissertation wurde am 17.04.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 30.10.2024 angenommen.

Matthias Fabian Keicher

Multimodal Deep Learning for Holistic Clinical Decision and Reasoning Support

Dissertation, Version 1.0

Technische Universität München

TUM School of Computation, Information and Technology

Lehrstuhl für Informatikanwendungen in der Medizin

Boltzmannstraße 3

85748 Garching bei München

Abstract

In clinical decision-making, medical doctors rely not only on a multitude of information about a patient, including patient history, vital signs, blood markers, and imaging data, but also on their extensive knowledge gained through formal education and experience with previously treated patients. To effectively support this complex process, clinical decision support systems should be able to integrate these different data modalities and incorporate relevant medical knowledge to make accurate diagnostic predictions. Furthermore, if these systems could provide insights into their reasoning process, they would not only support healthcare professionals more effectively in reasoning but also help build trust in the system's outputs, detect flaws in its reasoning process, and thereby enable continuous improvement. This thesis explores clinical decision support systems based on deep learning that integrate multimodal knowledge about a patient with formal and exemplar clinical knowledge while providing insight into their reasoning.

To learn from and adapt to the clinical decision-making process of medical doctors, we first detail the clinical reasoning process from a medical education and cognitive psychology perspective. We then give an overview of multimodal deep learning and how the heterogeneous modalities involved in the clinical decision-making process can be integrated into such models. The second part demonstrates how the exemplar knowledge about previously treated patients can be modeled with a population graph. Using multimodal patient data, we model the inter-patient relationships and the underlying patient characteristics. We leverage this advanced approach for toxin prediction at a poison control center, demonstrating that the system achieves superior performance compared to clinicians by additional fusion with textbook knowledge about symptoms. Next, we extend this approach with features extracted from computer tomography images and propose a novel method for multimodal population graph construction. We apply this to the outcome prediction of COVID-19 and show that the attention on relevant patients in the graph can be interpreted as mimicking the memory-retrieval reasoning of clinicians. In the third part, we explore how self-supervised pretraining on large amounts of unlabelled data can be used to extract structured knowledge from images and subsequently be used for transparent reasoning. In the first work, we predict structured radiology reporting elements for chest X-rays using only few annotated samples. We then introduce a novel zero-shot method, where instead of training, we make use of prior knowledge about disease manifestations and use this for transparent reasoning. Finally, we investigate extracting fine-grained semantic concepts from the neural activations of a deep learning model only trained on detecting vertebral fractures, assessing their radiological meaningfulness for potential utility in decision support. We conclude by providing an outlook on how the findings in this thesis could impact the rapidly growing field of multimodal large language models.

Zusammenfassung

Bei der klinischen Entscheidungsfindung stützen sich Ärzte nicht nur auf eine Vielzahl von Informationen über einen Patienten, wie Anamnese, Vitalwerte, Blutmarker und medizinische Bildgebung, sondern auch auf ihr umfangreiches Wissen, das sie durch formale Ausbildung und Erfahrung mit zuvor behandelten Patienten erworben haben. Um diesen komplexen Prozess wirksam zu unterstützen, sollten klinische Entscheidungshilfesysteme in der Lage sein, diese verschiedenen Datenmodalitäten zu integrieren und relevantes medizinisches Wissen einzubeziehen, um genaue diagnostische Vorhersagen zu treffen. Wenn diese Systeme darüber hinaus Einblicke in ihren Entscheidungsfindungsprozess gewähren könnten, würden sie nicht nur Mediziner gezielter bei der Entscheidungsfindung unterstützen, sondern auch dazu beitragen, Vertrauen in die Ergebnisse der Systeme aufzubauen, Fehlerquellen zu verstehen und so eine kontinuierliche Verbesserung zu ermöglichen. In dieser Arbeit werden klinische Entscheidungsunterstützungssysteme auf der Grundlage von Deep Learning erforscht, die multimodales Wissen über einen Patienten mit formalem und exemplarischem klinischem Wissen integrieren und gleichzeitig einen Einblick in ihre Entscheidungsfindung geben.

Um vom klinischen Entscheidungsprozess von Ärzten zu lernen und sich an diesen anzupassen, wird dieser zunächst aus der Perspektive der medizinischen Ausbildung und der kognitiven Psychologie beschrieben. Anschließend geben wir einen Überblick über multimodales Deep Learning und wie die heterogenen Modalitäten, die den klinischen Entscheidungsprozess beeinflussen, integriert werden können. Im zweiten Teil wird gezeigt, wie das exemplarische Wissen über bereits behandelte Patienten mit einem Populationsgraphen modelliert werden kann. Anhand multimodaler Patientendaten modellieren wir die Beziehungen zwischen den Patienten und ihre Merkmale. Wir setzen diesen Ansatz für die Toxinerkennung in einem Giftnotrufzentrum ein und zeigen, dass das System durch die zusätzliche Integration von Fachwissen über Symptome bessere Ergebnisse als Kliniker erzielt. Als Nächstes erweitern wir diesen Ansatz mit Merkmalen, die aus Computertomographie-Bildern extrahiert wurden. Wir wenden diese Methode auf die Vorhersage des Verlaufs von COVID-19 an und zeigen, dass die Aufmerksamkeit auf relevante Patienten im Graphen der Erinnerung eines Arztes an relevante Patienten entspricht. Im dritten Teil untersuchen wir, wie selbstüberwachtes Vortraining auf großen Mengen unstrukturierter Daten verwendet werden kann, um strukturiertes Wissen aus Bildern zu extrahieren und anschließend für nachvollziehbare Schlussfolgerungen zu nutzen. In unserer ersten Arbeit dazu bestimmen wir strukturierte radiologische Befundungselemente für Röntgenbilder der Lunge, indem wir nur wenige annotierte Datenpunkte nutzen. Anschließend stellen wir eine neuartige Zero-Shot-Methode vor, bei der wir Vorwissen über Manifestationen von Krankheiten in Radiologiebildern nutzen anstatt mit Daten zu trainieren. Im letzten Teil der Arbeit untersuchen wir die Extraktion feinkörniger semantischer Konzepte aus den neuronalen Aktivierungen eines Deep-Learning-Modells, das nur auf die Erkennung

von Wirbelkörperbrüchen trainiert wurde, und bewerten ihre radiologische Aussagekraft und ihren potenziellen Nutzen für die Entscheidungsunterstützung. Abschließend geben wir einen Ausblick darauf, wie sich die Ergebnisse dieser Arbeit auf das schnell wachsende Feld der multimodalen Large-Language-Modelle auswirken könnten.

Acknowledgments

I am deeply grateful to Nassir Navab for being an inspiration, always supporting me, and giving me the freedom to explore my research curiosity. I want to thank Thomas Wendler for giving me guidance, encouraging me, and building our lab's open and collaborative research environment. I am also grateful to the backbone of our CAMP chair, Martina Hilla, Ulrich Eck, Benjamin Bussam, Shahrooz Faghihroohi, Nikolas Brasch, and Zhongliang Jiang, for making our research possible.

Special thanks to all the friends at the chair I made throughout this journey, including but not limited to Tobias Czempiel, Hendrik Burwinkel, Magdalini Paschali, Farid Azampur, Walter Simson, Christine Eilers, Ashkan Khazar, Maria Tirindelli, and of course the vision-language team with Chantal Pelligrini, Kamilia Zaripova, David Bani-Harouni, and Ege Özsoy. I would also like to thank Paul Engstler, Matan Atad, Lukas Buess and all the other students who supported my work.

My research would not have been possible without the amazing clinical partners at the Klinikum rechts der Isar hospital. I am truly thankful to the Neuroradiology Department, in particular, Jan Kirschke and Benedikt Wiestler, Rickmer Braren from the Radiology Department, and Tobias Zellner and Florian Eyer at the Toxicology Department for providing a clinical perspective and taking the time to collaborate.

I am incredibly thankful to my parents, my sister, and my extended family for always being there for me and believing in me. Finally, my beloved wife and sons, thank you so much for your motivation, continuous support, and love.

Contents

I	Fundamentals of Multimodal Clinical Decision Support	1
1	Towards Clinical Reasoning Support Systems	3
2	Principals of Clinical Decision-Making	5
2.1	Clinical Decision-Making	5
2.1.1	The Differential Diagnosis Process	6
2.1.2	Uncertainty and Integrated Diagnostics	7
2.1.3	Clinical Reasoning	9
2.1.4	Clinical Education and Knowledge	10
2.2	Clinical Decision Support Systems	11
2.2.1	Knowledge Base vs. Machine Learning	11
2.2.2	Data Integration and Reasoning Support	11
2.2.3	The Need for Interpretability	12
3	Multimodal Deep Learning	13
3.1	Medical Modalities	13
3.2	Unimodal Representation Learning	15
3.2.1	Natural Language Processing	15
3.2.2	Spatial Data	17
3.2.3	Sequential Data	19
3.2.4	Tabular Data	19
3.3	Multimodal Representation Learning	20
3.3.1	Modality Fusion	21
3.3.2	Cross-Modal Translation	23
3.3.3	Self-supervised Representation Learning	24
II	Modelling Formal and Experiential Knowledge	27
4	Explicit Integration of Exemplar Knowledge in Deep Learning	29
4.1	Prototypical Networks	29
4.2	Retrieval of Exemplar Knowledge	30
4.3	Modelling Exemplar Knowledge with Graphs	30
5	Intoxication Prediction with Population Graphs and Medical Knowledge	33
5.1	Introduction	33
5.2	Methodology	34
5.2.1	Population Graph Processing	35
5.2.2	Integration of Formal Knowledge	36
5.3	Experimental setup	37

5.4	Results and Discussion	38
5.4.1	Ablative Testing and Baselines Comparison	38
5.4.2	Comparison with Clinicians	39
5.5	Conclusion	40
6	COVID-19 Outcome Prediction with Multimodal Population Graphs and Joint Pathology Segmentation	43
6.1	Introduction	43
6.2	Related Work	44
6.2.1	Integrating Imaging and Tabular Data	45
6.2.2	Graph Convolutional Networks	46
6.2.3	Multitask Learning	46
6.3	Method	47
6.3.1	Graph-based Image Processing	47
6.3.2	Segmentation, Image Features, and Radiomics	49
6.3.3	Multimodal Feature Fusion	50
6.3.4	Patient Outcome Prediction	50
6.4	Experiments	51
6.4.1	Multimodal COVID-19 Datasets	51
6.4.2	Implementation Details	55
6.4.3	Ablative Testing and Baselines	57
6.4.4	Metrics	58
6.5	Results and Discussion	59
6.5.1	Population Graph Construction	59
6.5.2	U-GAT Evaluation	59
6.5.3	Interpretability and Graph Attention	65
6.5.4	Challenges and Outlook	66
6.6	Conclusion	66
III	Cross-modal Extraction of Structured Knowledge	69
7	Structured and Unstructured Clinical Knowledge	71
7.1	Standardization and Structured Reporting	71
7.1.1	Structured Reporting in Deep Learning	72
7.1.2	Evaluating the Clinical Correctness of Reports	73
8	Contrastive Language-Image Pre-training for Structured Reporting of Chest X-rays	75
8.1	Introduction	75
8.1.1	Related Work	77
8.2	Method	78
8.2.1	Log-Sum-Exp Sign Loss	79
8.2.2	Contrastive Language-Image Pre-training	80
8.2.3	Cross-modal Similarity Metric	80
8.3	Experimental Setup	81
8.3.1	Structured Reporting Dataset	81
8.3.2	Implementation and Training Details	83
8.3.3	Few-shot Classification	85

8.4	Results	86
8.4.1	Ablation and Cardiomegaly Grading	86
8.4.2	Localization of Pathologies	86
8.5	Discussion	88
8.6	Conclusion	89
9	Zero-shot Classification of Chest X-rays with Deductive Reasoning on Radiological Findings	91
9.1	Introduction	91
9.2	Methodology	92
9.2.1	Model Overview	92
9.2.2	Prompt Engineering	94
9.3	Experiments and Results	94
9.3.1	Ablation Studies	96
9.3.2	Qualitative Results	99
9.4	Discussion	100
9.5	Conclusion	101
IV	Post-hoc Interpretation of Neural Networks	103
10	Interpretability	105
11	Explaining Vertebrae Fracture Detection with Semantic Concept Activations	107
11.1	Introduction	107
11.2	Related Work	108
11.2.1	Interpretability	109
11.3	Methodology	110
11.3.1	Vertebral Fracture Detection	110
11.3.2	Extraction of Semantic Concepts	110
11.3.3	Concept Correlation at Inference	111
11.4	Experimental Setup	111
11.5	Results and Discussion	112
11.5.1	Vertebral Fracture Detection	112
11.5.2	Clinical Evaluation of Semantic Concepts	113
11.5.3	Single-Inference Concept Visualization	114
11.6	Conclusion	115
V	Conclusion and Outlook	117
12	Conclusion	119
13	Outlook	121
VI	Appendix	123
A	Authored and Co-authored Publications	125
B	Abstracts of Publications not Discussed in this Thesis	129

Bibliography	137
List of Figures	157
List of Tables	161

Part I

Fundamentals of Multimodal Clinical
Decision Support

Towards Clinical Reasoning Support Systems

What if we could tap into the collective knowledge and experience of clinicians worldwide to provide the best care for every patient? As we see the world being transformed by advances in machine intelligence, the potential for the democratization and personalization of healthcare has never been greater. Imagine a future where an intelligent system analyzes all available patient characteristics in an instant and compares them against millions of previously examined patients and humankind's accumulated clinical knowledge, achieving higher scores than most clinicians. However, what if it cannot give us any insight into how it came to a diagnostic conclusion? Would we trust such a system? One of the reasons we trust medical doctors is because they involve us in the decision-making process and are able to explain their reasoning to us. But how does a clinical decision support system (CDSS) reason? Can it reason?

To understand the reasoning process of these systems, we must first understand how clinicians reason. Clinical decisions are complex and require not only clinical education and training of years but also hands-on experience with patients. Furthermore, clinicians need to integrate knowledge about a patient from many sources, including the patient's history, described symptoms, medical imaging, lab results, and more. How can we design models that form the foundation of such intelligent systems, enabling them to process and reason over this complex, multimodal data?

This thesis is built on the belief that intelligent systems will not replace physicians but rather support them in aggregating relevant information from complex heterogeneous data and help them reason over these findings. Therefore, beyond making predictions, clinical decision support systems should provide insights into their reasoning. This will support clinicians' reasoning, build trust, and make the systems' potential faults and biases more transparent. Ultimately, such clinical *reasoning* support systems will allow clinicians to pay more attention to patients and provide better personalization of medicine.

The objective of this thesis is to understand the clinical reasoning process and, based on this, investigate how decision support systems can be designed to integrate multimodal knowledge about individual patients with previous patients and formal textbook knowledge while giving insights into their reasoning.

In the first part of this thesis, we first lay the theoretical foundations of the medical decision-making process (Chapter 2) from both a cognitive and medical education perspective to understand the reasoning and the type of clinical knowledge involved. Next, we discuss existing clinical decision-support systems that can facilitate this process. Finally, we give an overview of multimodal deep learning (Chapter 3) and how heterogeneous modalities used in clinical decision-making can be integrated into deep learning models.

The next part investigates how experiential knowledge (Chapter 4) about previous patients can be modeled with multimodal patient population graphs. For the prediction of intoxication

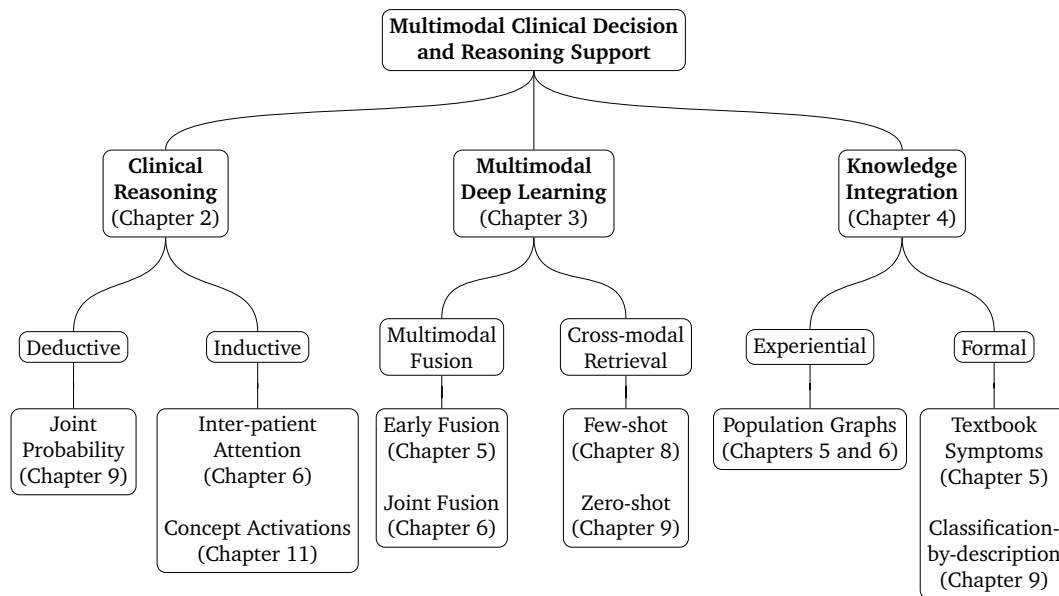


Fig. 1.1. Overview of the contributions of this thesis.

at a poison control center (Chapter 5), we integrate this exemplar knowledge with literature knowledge about occurring symptoms. Building on this work, we investigate further how images can be integrated with other clinical information about the patient and put into perspective with other patients in the graph for the outcome prediction of COVID-19 (Chapter 6). In the third part of the thesis, we investigate how we can extract patient-specific radiological knowledge using large amounts of unlabelled pairs of radiology reports and images. First, we use these multimodal patient representations to predict structured reports for a given chest X-ray using only a small number of annotated samples (Chapter 8). To give more insight into the reasoning process of this method, we then extend it to a classification-by-description approach (Chapter 9). Here, we first compile descriptions of the manifestations of suspected diseases and then match them against a given image, mimicking the deductive reasoning process.

The final part is dedicated to exploring the interpretation of deep learning models with no intrinsic interpretability. It tries to answer the question of whether we can retrospectively shed light on the reasoning behind the automatic detection of vertebral body fractures in CT images (Chapter 11).

Figure 1.1 provides an overview of the three pillars of contributions in this thesis. We investigate the similarities between the cognitive process of clinicians and deep learning models and introduce distinct reasoning concepts that correspond both to the intuitive inductive being a natural fit to data-driven models and deductive reasoning. Towards the holistic integration of heterogeneous patient data in deep learning, we propose fusion strategies on different levels of abstraction and explore the extraction of patient-specific knowledge across modalities. Taking further inspiration from the clinical mind, we then propose methods integrating both the knowledge about individual patients that reflect the experience of medical doctors and the formal knowledge they acquire in their education.

Principals of Clinical Decision-Making

Contents

2.1	Clinical Decision-Making	5
2.1.1	The Differential Diagnosis Process	6
2.1.2	Uncertainty and Integrated Diagnostics	7
2.1.3	Clinical Reasoning	9
2.1.4	Clinical Education and Knowledge	10
2.2	Clinical Decision Support Systems	11
2.2.1	Knowledge Base vs. Machine Learning	11
2.2.2	Data Integration and Reasoning Support	11
2.2.3	The Need for Interpretability	12

Making clinical decisions is hard and requires a lifetime to master. Clinicians often face high-stakes decisions with limited time and resources, making choices that can impact life or death based on only a fraction of the necessary information for well-informed decision-making. To deal with this uncertainty, they run through years of formal education only to be able to start practical training in a specialized field that takes many years to complete. The first part of this chapter gives an overview of the theoretical foundation and research on the clinical decision-making process and reasoning as well as the knowledge and education of this knowledge that is required to perform these with the best outcome for the patient. The second part discusses how decision-support systems can aid clinicians and what systems have been developed and used in the past.

2.1 Clinical Decision-Making

To effectively model clinical decision support systems, it is crucial to understand how clinicians arrive at their diagnostic and treatment decisions. The clinical decision-making process can be defined as the process of analyzing a patient's status to decide on the ideal treatment and, therefore, optimal patient outcome [212]. The initial step in the clinical decision-making process is accurately diagnosing the underlying disease causing the patient's symptoms. This is essential to determine the most effective treatment to address the root cause of a patient's health issues rather than merely alleviating the symptoms.

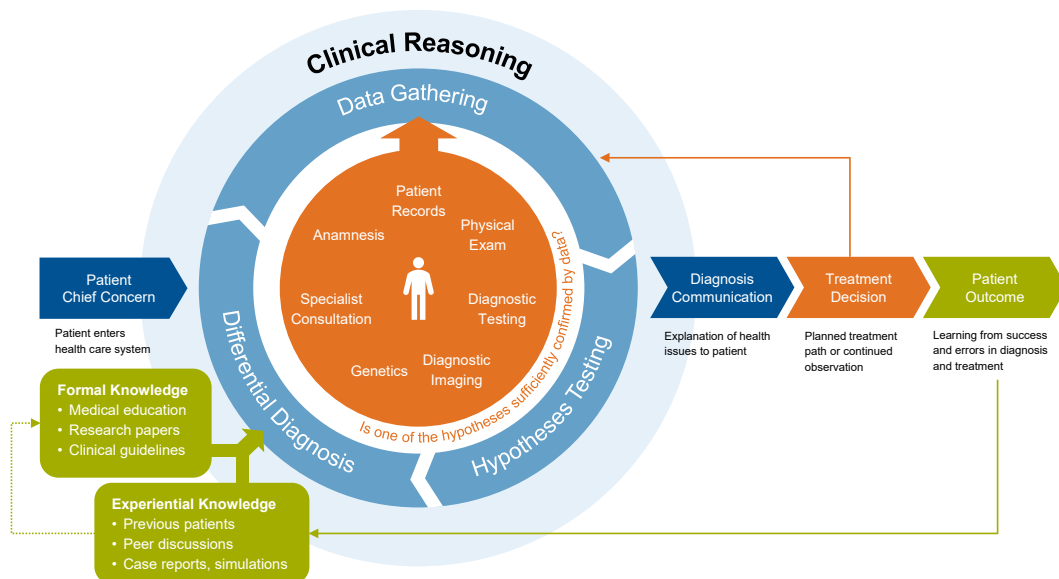


Fig. 2.1. The clinical decision-making process. The process begins when a patient presents with a chief concern. The clinician generates a differential diagnosis based on the patient's symptoms and available data. Next, the differential is narrowed to the most likely diagnosis through an iterative clinical reasoning process. This diagnosis is then communicated to the patient, and a treatment plan is established. The clinician relies on formal and experiential knowledge throughout this process, continuously improving through practice and education. This process is adapted from the diagnostic process outlined by Balogh et al. [53] incorporating the cyclic reasoning process of differential diagnosis described by Sox et al. [219].

2.1.1 The Differential Diagnosis Process

The diagnostic decision-making process (Figure 2.1), as described by Balogh et al. [53], is typically initiated when a patient experiences a health problem and presents their chief complaint to a healthcare professional. Based on this self-reported description of symptoms and health concerns, the clinician forms an initial set of hypotheses regarding potential diagnoses, known as differential diagnosis [219].

At the heart of the process is the iterative clinical reasoning process, which narrows the list of possible diagnoses by testing and refining hypotheses with newly gathered patient information. This patient data is acquired based on the hypothesis to be tested and includes patient interviews, physical exams, patient records, and diagnostic tests like blood tests or diagnostic imaging [219]. The clinician analyzes and integrates each new piece of information to determine whether it supports or contradicts the different hypothesized diagnoses until a most likely diagnosis or set thereof is identified with sufficient evidence. Once a working diagnosis has been established, the clinician communicates the findings to the patient and discusses the available treatment options. The patient's response to the selected treatment can provide further information to refine or modify the differential diagnosis, triggering the clinical reasoning process again, if required. Throughout the diagnostic process, clinicians rely on a combination of formal knowledge acquired through medical education and experiential knowledge gained from previous clinical experiences that are further discussed in Section 2.1.4 [174]. In addition to the subconscious building of clinical expertise, systematic capturing and analysis of patient outcomes is essential for improving evidence-based medicine and refining diagnostic and treatment guidelines [24, 64].

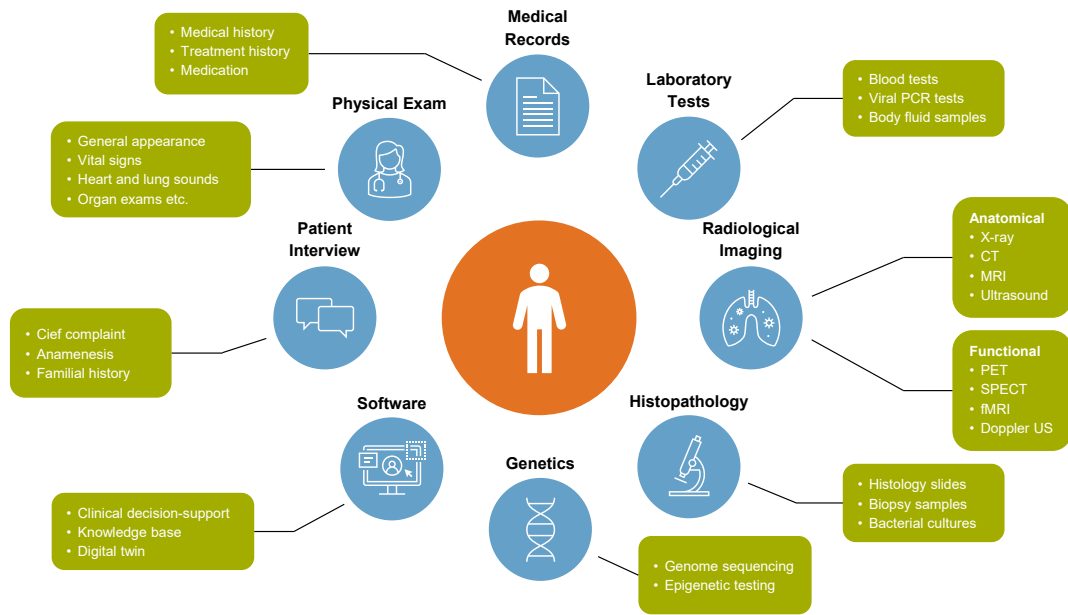


Fig. 2.2. Integrated diagnostics. To find the best treatment for a patient, clinicians need to form a holistic representation of the patient to confirm or rule out hypotheses in the differential diagnosis. Adapted from [1, 16, 53, 161]

2.1.2 Uncertainty and Integrated Diagnostics

In an ideal world, each diagnosis would be associated with a fixed set of clinical findings, where their presence or absence would provide clear evidence of a specific disease, making the diagnostic process straightforward and eliminating the need for complex reasoning. However, the reality of clinical practice is far from this idealized scenario [219].

Every clinical finding merely indicates a probability of various diseases, and symptoms manifest differently in each patient. Moreover, each finding and diagnostic test is associated with inherent uncertainty, as well as the potential for diagnostic and human error [162]. Therefore, clinicians must consider the sensitivity and specificity of a given diagnostic test when interpreting results. The clinical decision-making process can be modeled as reducing uncertainty in the differential diagnosis, aiming to identify the diagnosis with the highest probability for a given patient [206] (see Figure 2.3). Clinicians naturally think in probabilities, and the impact of a diagnostic test on the probability of a hypothesized diagnosis can be formalized using Bayes' Theorem, as shown in Equation 2.1 [219]. Given a disease probability prior to the test $P(D)$, the posterior probability after testing $P(D|T)$ can be calculated using the probability of a given test result with the disease present $P(T|D)$ and the general probability of said test result $P(T)$.

$$P(D|T) = \frac{P(T|D) \times P(D)}{P(T)} \quad (2.1)$$

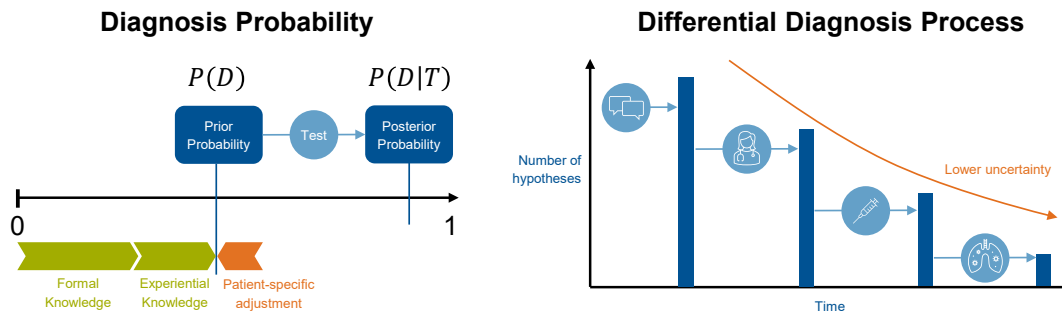


Fig. 2.3. Uncertainty reduction during the differential diagnosis process. Adapted from Sox et al. [219]

A more practical variation of this equation for the probability of a disease being present (D_+) given a positive test result (T_+) using the sensitivity ($P(T_+|D_+)$) and specificity ($P(T_-|D_-)$) of a diagnostic test can be represented as [219]:

$$P(D_+|T_+) = \frac{P(D_+) \times \text{Sensitivity}}{P(D_+) \times \text{Sensitivity} + (1 - P(D_+)) \times (1 - \text{Specificity})} \quad (2.2)$$

This equation highlights the complexity of integrating a diagnostic test in the differential diagnosis process, as clinicians must consider the test's sensitivity, specificity, and the pre-test probability of the disease to accurately interpret the results and update their diagnostic hypotheses. As shown in Figure 2.3 a doctor might form an initial set of hypotheses based on his formal knowledge and experience and assign each considered diagnosis a probability based on the known characteristics of the patient's health status. By adding more information like diagnostic testing, the doctor can rule out diagnoses and reduce the overall uncertainty of his differential.

In theory, doctors should consider all relevant information when making diagnostic decisions. However, they may not have access to all necessary data or must weigh the cost-benefit of a diagnostic test in the context of treatment urgency and healthcare system costs.

To effectively diagnose and treat patients, clinicians must integrate information from various sources, as illustrated in Figure 2.2. To gather important information about a patient's health, medical professionals start with basic data found in the patient's medical records. This data includes the patient's demographics, medical history, treatment history, and medication. They may also conduct patient interviews to gather more information about their medical history, including their family history, and perform physical examinations and measurements like blood pressure to get a complete picture of their health. Diagnostic tests can be carried out to determine the cause of a medical condition. These include blood tests to measure hormones and inflammation markers, viral PCR tests, and genomic analyses. [1, 16, 53, 161]

Diagnostic radiology and nuclear medicine is an essential part of the diagnosis process. It involves obtaining and interpreting medical images to confirm or rule out suspected diseases or clinical findings. It also provides consultative information to help with clinical decision-making within the broader context of patient care [136]. Radiological imaging techniques such as X-ray, CT, MRI, and ultrasound can also be used to identify anatomical abnormalities.

Furthermore, functional and nuclear imaging techniques like scintigraphy, PET, and SPECT can provide additional physiological information. Other diagnostic tests, including ECG, EEG, and endoscopy, can help diagnose medical conditions. Histopathology results from histology slides, biopsy findings, and bacterial growth assessments can provide further insights into the cause of a condition. [1, 16, 161]

Diagnosis and outcome prediction is particularly challenging in oncology [161], where a multi-disciplinary approach involving a tumor board is required to integrate all available patient information and expert opinions from various clinical fields, such as pathology and radiology.

2.1.3 Clinical Reasoning

How do clinicians form a diagnosis for a given patient to be able to assess the best treatment? In the literature, the cognitive task of clinical reasoning has been studied from various academic perspectives, such as medical education and cognitive psychology. The most widely accepted mental model is that clinicians employ both intuitive and analytical thinking in their decision-making processes [133, 212]. While experienced clinicians can quickly recognize a familiar pattern of symptoms in a new patient based on previous encounters, novices rely on a more systematic way of building a hypothesis using factual knowledge and then looking for supporting data in their mental patient model [212].

Dual Process Theory

This model is formalized in the dual process theory, popularized by "Thinking, Fast and Slow" by Daniel Kahneman [119]. It classifies decision-making into two types of cognitive processes: System 1, the intuitive approach, and System 2, the analytical mind. System 1 corresponds to effortless and fast decision-making based on heuristics and intuition formed by experience. While this subconscious, "gut-feeling"-based decision-making is highly efficient in detecting well-known symptom patterns and dealing with uncertainty, it can be prone to cognitive bias, like sticking with an initial diagnosis even though contradicting evidence may arise. [55]

In contrast, System 2 represents slow and thorough decision-making involving systematic analysis, resulting in a low error rate if correct knowledge is used and all required patient information is available. It is important to note that these two modes of thinking are not exclusive but rather the two extremes of a continuum that a practitioner can jump between at any time. An experienced doctor might divert from intuitive to analytical thinking when faced with a rare disease or unusual combination of symptoms. At the same time, a novice requires intuition to develop an initial hypothesis for a differential diagnosis. Through the lens of cognitive psychology, System 1 is associated with inductive reasoning, which uses pattern recognition to generate hypotheses in a bottom-up manner. On the contrary, System 2 aligns with deductive reasoning, which systematically tests hypotheses against data in a top-down approach. [55, 133]

Memory-based Pattern Recognition

Memory-based theories for reasoning provide a different perspective on decision-making at the intuitive end of the spectrum (System 1) by not focusing on the cognitive process but on

what knowledge is retrieved from memory to recognize patterns instead. Two main types of mental models for categorizing diseases by retrieving experiential knowledge from memory are exemplar and prototype knowledge. Exemplar models suggest that the most relevant and similar individual experiences about specific patients are retrieved from memory to match patterns when diagnosing a new case. In prototype models, on the other hand, the memory retrieves an abstract, average representation of past experiences associated with a particular disease or finding. Both of these models use experiential knowledge, which is mainly formed by experiences of past patient encounters, diagnoses, and treatments. [133, 165]

There is also evidence for highly effective knowledge structures of information about diseases in the memory of doctors: illness scripts and semantic qualifiers. Illness scripts are mental templates that link relevant information about a disease, such as enabling conditions, causes, and consequences, into a coherent narrative representation. They allow physicians to efficiently activate diagnostic hypotheses based on the patient's presenting signs and symptoms. A simpler yet effective knowledge representation is semantic qualifiers, which are abstract, dichotomous descriptors that help to quickly decide between competing diagnostic hypotheses, e.g., acute vs. chronic, progressive vs. improving, and unilateral vs. bilateral. [97, 133]

2.1.4 Clinical Education and Knowledge

An impactful application of clinical reasoning theory is understanding where diagnostic errors originate from and how clinical education can be improved with these insights. Norman et al. [175] explore whether diagnostic errors primarily originate from System 1 or System 2 thinking. The authors conclude that although mistakes can stem from both systems, the most significant potential for reducing diagnostic errors lies in enhancing clinicians' knowledge.

Formal Knowledge

In the early stages of their studies, medical students primarily acquire basic science knowledge, which provides the foundation for understanding human anatomy and physiology. After understanding the mechanisms underlying various diseases (etiology), they finally learn how diseases can be treated. Therefore, the formal network of knowledge required to make clinical decisions can be categorized into diagnostic knowledge, etiological knowledge, and treatment knowledge. At this stage, clinicians have limited practical experience and rely on deductive reasoning to confirm a differential diagnosis by analyzing the patient's characteristics involving detailed biomedical concepts, which are organized in knowledge networks. [97, 212]

The acquisition of formal knowledge does not end with the completion of a clinician's formal education; rather, it is an ongoing process that is particularly important when new insights into diseases are gained or when evidence-based guidelines are updated based on the latest research and expert consensus [64].

Experiential Knowledge

By applying this formal knowledge in clinical practice, students begin to encapsulate biomedical concepts into clinical knowledge, enabling them to draw direct conclusions without explicitly referring to the fundamentals. This process, known as knowledge encapsulation, involves clustering related concepts together, allowing students to directly connect patient findings and clinical hypotheses or diagnoses [97]. With more experience, the knowledge

organization transitions from a network-like structure to a more efficient and integrated format known as illness scripts as described in Section 2.1.3. As clinicians gain more experience and refine their illness scripts, they become increasingly adept at pattern recognition and intuitive decision-making, relying more on System 1 thinking. However, when faced with complex or unfamiliar cases, experienced clinicians can still unfold their encapsulated biomedical knowledge and employ System 2 to resort to analytical problem solving [97, 212]. Monteiro et al. [165] argue that only learning and applying knowledge - making mistakes on the way - can turn novices into experts excelling at clinical decision-making. They propose that experiential knowledge is so crucial that medical education should use more simulation-based learning environments to compensate for a lack of experience in particular with rare diseases.

2.2 Clinical Decision Support Systems

Clinical decision support systems (CDSS) are designed to assist clinicians in the decision-making process, as explained in the previous chapter. These computer-aided systems aim to support clinicians in making accurate and cost-effective decisions, ultimately improving patient outcomes by assisting in cognitively demanding tasks such as probabilistic reasoning and pattern recognition in complex data [234].

2.2.1 Knowledge Base vs. Machine Learning

There are two main types of CDSS: knowledge-based and machine learning-based [221]. Knowledge-based systems, which reflect the analytic (System 2) decision-making process, rely on a comprehensive knowledge base that encapsulates formal domain knowledge, such as clinical guidelines, evidence-based medical insights, and research findings. This knowledge is structured in decision trees or knowledge graphs, enabling the system to imitate the analytical reasoning process of human experts by following decision rules. In contrast, machine learning-based CDSSs align with medical doctors' intuitive (System 1) reasoning and are data-driven and learn decision-making rather than being programmed with expert knowledge. While machine learning-based systems have shown impressive performance, particularly in medical image analysis [196], they often lack inherent interpretability, which limits trust in the systems and insights into the reasoning of these models, in contrast to knowledge-based systems.

2.2.2 Data Integration and Reasoning Support

A significant benefit of CDSS is their ability to integrate various data sources, including electronic health records (EHRs) and biometric monitoring, to provide comprehensive support and highlight relevant characteristics [221]. In the future, interactive decision support systems could retrieve all relevant information to support diagnosis from the various distributed data sources within a hospital's electronic information system and even personal wearables and other smart sensors [16]. By integrating these diverse data sources and leveraging advanced analytical techniques, such systems have the potential to enhance diagnostic accuracy and support clinicians in providing optimal patient care.

Recent advancements in Language Models (LLMs) and other foundation models suggest that future CDSS will be capable of dealing with both structured and unstructured health data, providing both knowledge-driven and learned decision support [166]. This development could revolutionize the way CDSS operates, allowing for more comprehensive and intuitive support for clinicians. Following the vision of van Baalen et al. [234], future CDSS could act as clinical reasoning support systems that support the reasoning process of clinicians rather than making decisions for them. By augmenting the clinician's decision-making process and providing relevant insights and recommendations, CRSS could enhance the overall quality of care while maintaining the clinician's autonomy and expertise.

2.2.3 The Need for Interpretability

Do we need clinical decision-support systems to be interpretable? Rule-based systems offer high levels of interpretability but are often outperformed by data-driven models that lack inherent transparency in their decision-making processes. This raises the question of whether predictive performance should always be prioritized over interpretability, assuming there is a trade-off between the two. As illustrated in the clinical decision-making process (Figure 2.1), communicating and explaining reasoning to patients is essential since actively involving them can improve clinical outcomes [181]. Experienced clinicians are expected to provide well-reasoned, analytical explanations of their assessments to patients, even when making decisions based on intuition and heuristics alone. The potential trade-off between interpretability and performance in decision support systems is an active area of research and discussion within the scientific community. However, there is a growing consensus that offering transparency in the decision-making can build trust in (semi-)automated systems and help to detect and correct systematic errors during their development [205].



- Expert clinicians excel at intuitive decision-making, while novices rely on analytical thinking.
- Experts revert to analytical thinking for difficult or rare cases, considering all available patient data.
- Clinical reasoning involves different knowledge types (formal vs. experiential) and structures (knowledge networks vs. illness scripts).
- Machine learning mimics intuitive decision-making and excels with enough data, but analytical models like decision trees offer higher interpretability.

Multimodal Deep Learning

Contents

3.1	Medical Modalities	13
3.2	Unimodal Representation Learning	15
3.2.1	Natural Language Processing	15
3.2.2	Spatial Data	17
3.2.3	Sequential Data	19
3.2.4	Tabular Data	19
3.3	Multimodal Representation Learning	20
3.3.1	Modality Fusion	21
3.3.2	Cross-Modal Translation	23
3.3.3	Self-supervised Representation Learning	24

In the previous chapter, we discussed the importance of integrating all available patient data to construct a comprehensive patient model in diagnostic and prognostic decision-making. Unlike rule-based systems, as outlined in Section 2.2, deep learning models learn to solve tasks directly from data rather than relying on feature engineering and pre-programmed instructions based on existing knowledge. A fundamental aspect of deep learning is representation learning [19], which involves learning the most informative features or representations from input data to optimize performance on one or more downstream tasks, such as classifying diseases or segmenting tumors. These representations can be learned with or without a task-specific supervision signal, known as supervised and unsupervised learning [19]. Additionally, self-supervised learning, where the supervision signal is derived from the input data itself, is another approach that will be further explored in Section 3.3.3. Multimodal deep learning is concerned with integrating representations from multiple modalities and has been extensively researched for various applications, including sensor fusion in autonomous vehicles, human activity recognition, video classification, and medicine [105, 197]. This chapter will delve into various clinical data modalities that can be processed with deep learning (Section 3.1), the methods for encoding these modalities for downstream tasks (Section 3.2), and the integration of multiple modalities to generate multimodal patient representations (Section 3.3).

3.1 Medical Modalities

The need for integrating different modalities (Section 2.1.2) for clinical decision-making can be highlighted by the examples of detecting hyperthyroidism or pancreatic cancer. To diagnose hyperthyroidism a medical doctor needs to consider patient-reported symptoms (fatigue, overweight, low activity, depression, etc.), vital signs (blood pressure and heart rate, etc.), thyroid ultrasound imaging, and thyroid-related blood tests [16]. Similarly, patient-reported

Tab. 3.1. Overview of data modalities, data types, and common deep learning architectures used in clinical decision support.

Data Modality	Data Type	Deep Learning Models
Text Data (Section 3.2.1)	Medical history	Transformer-based models (BERT, GPT)
	Treatment history	
	Patient medical records	Sequential models
	Patient interviews	
	Family history	
	Physical examination notes	
	Radiology reports	
Histopathology reports		
Imaging Data (Section 3.2.2)	Radiological imaging (X-ray, CT, MRI, Ultrasound)	Convolutional Neural Networks (CNNs) Vision Transformers (ViTs)
	Functional and nuclear imaging (Scintigraphy, PET, SPECT)	
	Optical coherence tomography (OCT)	
	Fundus imaging	
	Histopathology slides	
Time-Series and Sequential Data (Section 3.2.3)	Electrocardiogram (ECG)	Recurrent Neural Networks (RNNs)
	Electroencephalogram (EEG)	Long Short-Term Memory (LSTM)
	Sound recordings	Gated Recurrent Units (GRUs)
	Surgical videos	Temporal Convolutional Networks (TCNs)
	Genome sequences	Transformer-based models Hybrid models (CNN+TCN)
Tabular Data and Structured Data (Section 3.2.4)	Treatment codes (ICPM)	Multilayer Perceptrons (MLPs)
	Diagnosis codes (ICD)	Convolutional Neural Networks (CNNs)
	Patient demographics	Transformer-based models
	Vital signs	
	Blood test results	
	Viral PCR test results	
	Genomic analyses	
	Standardized Reports	

symptoms may trigger the suspicion of pancreatic cancer, and subsequently, the diagnosis needs to be confirmed through various imaging modalities (e.g., ultrasound and PET/MR) and laboratory results like PSA levels [16].

In addition to the typical challenges associated with curating biomedical data for deep learning, such as data privacy concerns and data imbalance, multimodal medical data brings an additional set of challenges, like data heterogeneity [1]. The dimensionality of the data can vary greatly, and dense data, e.g., CT images, may need to be integrated with sparse data, e.g., EHR data. Furthermore, missing information, such as unavailable laboratory test results for some patients, can complicate the fusion of data and necessitate the use of methods like mean imputation to handle the missing values.

Table 3.1 summarizes the wide range of modalities employed in clinical decision-making and their corresponding deep learning methods. Most modalities can be categorized or converted into text data (Section 3.2.1), imaging data (Section 3.2.2), sequential data (Section 3.2.3), or tabular data (Section 3.2.4).

Given the prevalence and large-scale availability of chest X-rays, a significant portion of research on multimodal deep learning, particularly vision-language models, has focused on chest X-ray datasets such as MIMIC-CXR [117] and CheXpert [108]. Several large-scale multimodal datasets featuring paired CT images and radiology reports have recently been released, which will also allow the training of foundational multimodal models for this imaging modality in the future [89, 104, 225]. Table 3.2 includes a selection of multimodal datasets for research purposes. Furthermore, research databases like the UK Biobank and The Cancer Imaging Archive offer extensive collections of multimodal patient data to support advancing the field of multimodal deep learning in healthcare.

3.2 Unimodal Representation Learning

To lay the foundation for understanding multimodal representation learning, it is essential to first examine how deep learning models learn representations from individual modalities. In the following sections, we will explore the key approaches and architectures employed in learning representations from text (Section 3.2.1), images and spatial data (Section 3.2.2), time-series and sequential data (Section 3.2.3), and tabular data (Section 3.2.4).

3.2.1 Natural Language Processing

Natural language processing (NLP) plays a crucial role in clinical decision-making, as text and speech are the primary means of documenting and communicating knowledge in healthcare. NLP models can extract relevant information from unstructured medical records, such as patient history, physical examination notes, and discharge summaries. The learned representations can also be applied to various tasks, including classifying structured information like diagnosis and treatment codes and selecting cohorts from clinical notes. Additionally, generative models can produce human-like text for applications such as automated report generation and question-answering systems.

For processing with deep learning models, text is modeled as sequential data by tokenizing words and subwords to a series of discrete token embeddings, which are dense vector representations of each token based on a fixed vocabulary. Therefore, the methods described for time-series analysis in Section 3.2.3, such as Recurrent Neural Networks (RNNs), have been applied extensively for NLP. However, the introduction of Transformer [236], and its self-attention mechanism has revolutionized the NLP field. Proposed initially for machine translation, self-attention allows models to attend to different parts of the input sequence when encoding each element, effectively weighing the importance of words based on their relevance to the task. This enables Transformers to capture long-range dependencies and contextual relationships more effectively than RNNs, while also allowing for parallel computation. In contrast, RNNs have to keep internal representations while processing the tokens sequentially. Positional embeddings are also added to each token to provide the Transformer with information about its position within the sequence.

Two of the most impactful architectures built on Transformers are the Generative Pre-trained Transformer (GPT) [195] and Bidirectional Encoder Representations from Transformers

Tab. 3.2. Overview of multimodal datasets for various clinical applications.

Dataset	Modalities	Description	Samples
ABIDE [63]	R-fMRI Structural MRI Phenotypic data	fMRI datasets from individuals with Autism spectrum disorders (ASDs) and age-matched control group. Includes data from 17 international sites.	1112 total 533 ASD 579 controls
TADPOLE (ADNI) [188]	PET-MR images Radiomics CSF biomarkers Genetic markers Demographics	The TADPOLE challenge, a subset of the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset, includes data aimed at predicting the progression of Alzheimer's Disease.	229 normal 398 with MCI 192 with AD 819 patients
MIMIC-CXR [117]	Chest X-rays Radiology reports	A comprehensive dataset of radiographic studies of 65,379 patients collected at Beth Israel Deaconess Medical Center. Includes multiple views, accompanied by semi-structured free-text radiology reports.	377,110 images 227,835 studies 65,379 patients
IU X-ray [60]	Chest X-rays Radiology reports Structured findings	Open-i collection of chest X-rays with corresponding reports from the Indiana University hospital network. Radiological findings and diagnoses are also encoded with Medical Subject Heading (MeSH) and RadLex encodings.	7,470 studies
CheXpert [108]	Chest X-rays Radiology reports Structured findings	The CheXpert dataset is a large public dataset containing chest radiographs from 65,240 patients. It includes labels for 14 common thoracic pathologies and observations extracted from radiology reports considering uncertainty.	224,316 X-rays 65,240 patients
RadFusion [273]	CT images Structured EHRs	A multimodal pulmonary embolism database with high-quality CT images and longitudinal patient EHR data, including demographics, vitals, medications, ICD codes, and lab tests.	1,837 studies 1,794 patients
CTRG-Brain-263K [225]	Brain CT Radiology reports	A dataset containing 263,670 brain CT scans with diagnostic reports. Aimed at generating medical reports for a series of radiological images of the brain.	263,670 studies
CTRG-Chest-548K [225]	Chest CT Radiology reports	Similar to the CTRG-Brain-263K dataset but focused on the chest area for detailed pathology assessment. It consists of 548,696 chest CT scans with diagnostic reports.	548,696 studies
iCTCF [170]	Chest CT Clinical features	The iCTCF (integrative CT images and Clinical Features for COVID-19) dataset comprises data from 1,521 patients. It includes chest CT images and 130 clinical features from biochemical and cellular analyses of blood and urine samples.	19,685 CT slices 1,521 patients
CT-RATE [89]	Chest CT Radiology reports Structured findings	The CT-RATE dataset comprises 3D chest CT volumes from 21,304 patients, paired with corresponding radiology reports.	50,188 volumes 25,692 patients
INSPECT [104]	CTPA images Radiology reports Structured EHRs	CTPA studies from 19,402 patients at risk for pulmonary embolism. Includes de-identified CT images, report impressions, and longitudinal EHRs (diagnoses, procedures, vitals, meds).	23,248 studies 19,402 patients

(BERT) [61]. Both models have demonstrated that pre-training on large amounts of unlabeled data allows Transformers to learn rich representations and powerful generative capabilities. GPT employs an autoregressive language modeling objective, predicting the next word in a sequence to develop complex language understanding and generation abilities when trained at scale. In contrast, BERT utilizes a masked language modeling objective, where the model predicts randomly masked words in a sequence, enabling it to learn bidirectional contexts and create robust representations suitable for various downstream tasks. As the pre-training corpus defines the knowledge embedded in these language models, a series of domain-specific models have been proposed for biomedical applications, such as BioGPT [154] and CXR-BERT [25]. The global embeddings of the input text, i.e., the whole token sequence, can be extracted from a dedicated classification token or by averaging the embedding of the individual tokens in the sequence. This representation can then be combined with representations from other modalities, as discussed in Section 3.3.1, to support multimodal clinical decision support systems.

3.2.2 Spatial Data

Understanding the content of medical images is crucial for incorporating radiological findings into the clinical decision-support process. Deep learning has been applied to various medical imaging modalities, such as radiology, histopathology, dermatology, and ophthalmology [72]. The supervision signal for these models can originate from either image-level tasks, like classifying or regressing clinical metrics, or pixel-level objectives, such as semantically segmenting tissue types in radiology images or detecting cancerous tissue in pathology slides using bounding boxes [139, 196]. At the heart of these architectures lies an image encoder that transforms the input into an abstracted latent representation, typically with a reduced spatial resolution [19].

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) achieve this by applying learned filters to local image patches, enabling pattern detection across different spatial locations, as illustrated in Figure 3.1. Pooling layers in CNNs merge semantically similar features through local maximum or average operations, introducing invariance to minor shifts and distortions while reducing the representation's spatial dimensions. By iteratively applying convolutional and pooling operations, CNNs learn multi-level representations, composing lower-level features into higher-level ones. This process mimics the hierarchical processing in the human visual cortex system, where information progresses from edges to motifs, parts, and ultimately, objects, forming a hierarchy of increasingly abstract features [139]. For classification tasks, the resulting feature map is typically pooled into a latent vector, removing its spatial resolution, and then forwarded to a single-layer neural network with the latent dimension as input and the number of classes as output, as shown in Figure 3.1. This representation can also be a useful image representation for joint fusion approaches, as described in Section 3.3.1, to integrate with other modalities at the same abstraction level for further processing.

Well-established CNN architectures include ResNet [92] and DenseNet [102]. For pixel-wise predictions, the encoder is followed by an upsampling decoder path involving upscaling or deconvolutions. The de facto standard architecture for pixel-wise predictions is based on

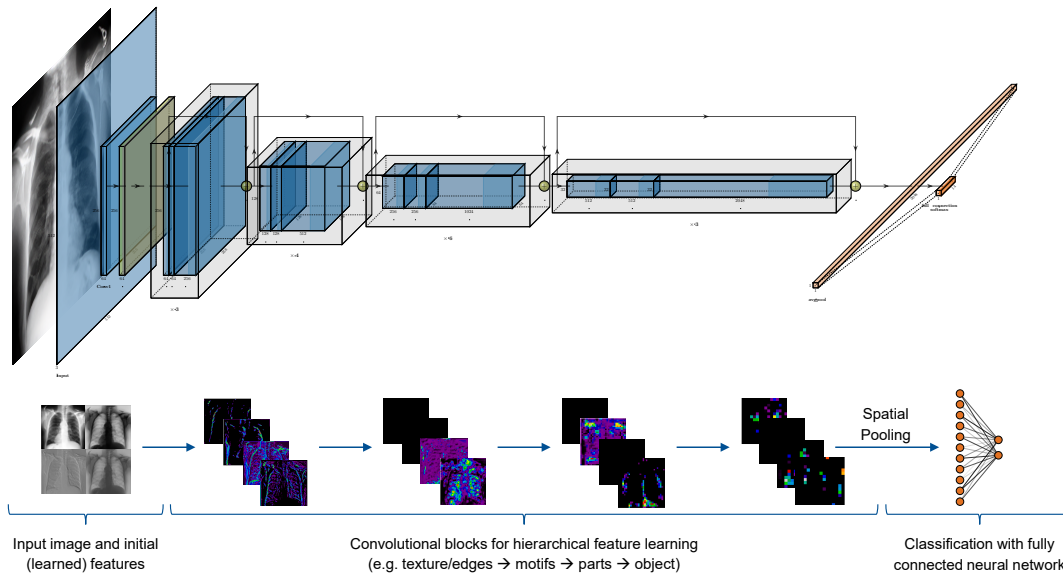


Fig. 3.1. ResNet50 is a well-known Convolutional Neural Network commonly used for encoding and classifying medical images proposed by He et al. [92]. Here, the classification of 14 findings is visualized with a synthetic X-ray.

U-Net [199] and its optimized variants [109], which utilize skip connections in the decoder to integrate input features at the relevant abstraction level for upsampling. CNNs are not limited to two dimensions but can be adapted to one dimension (see Section 3.2.3) or three dimensions for volumetric images like MRI [163] and even four dimensions for dynamic volumetric imaging.

Vision Transformer

Inspired by the success of Transformers in Natural Language Processing (NLP), Dosovitskiy et al. [65] introduced Vision Transformers (ViTs) that treat image patches as a sequence of tokens, similar to word embeddings in a sentence. In the original ViT, patches are embedded using a non-convolutional neural network and encoded with two-dimensional spatial positional encoding to preserve their spatial relationships. A Transformer Encoder then processes the patch embeddings, and the global image representation can be extracted using a classification token like BERT or by averaging the resulting patch embeddings.

For multimodal data fusion, either the global image embedding or the sequence of individual image patch tokens can be utilized for localized representations, as discussed in Section 3.3.1. Since ViTs need vast amounts of training data, they usually rely on transfer learning from large-scale 2D pretraining, which limits their effective use on small 3D datasets commonly found in the medical domain. To address this, our work Video-CT-MAE [28] demonstrates that ViTs can also be initialized with weights from models pretrained on natural videos. Despite efforts to facilitate ViT training in low data regimes [34], CNNs remain a robust choice for vision encoders in medical image analysis.

3.2.3 Sequential Data

Time-series and sequential data play a vital role in clinical decision-making and have been researched with deep learning methods, with applications such as activity monitoring using Electrocardiography (ECG) [148] and epileptic seizure prediction using Electroencephalography (EEG) [112]. Genome sequences can also be modeled as sequential data and processed with Transformer-based models like DNABERT [113], showcasing their effectiveness in analyzing DNA. Moreover, the temporal changes of static patient data, such as blood test results or sequential imaging data, as demonstrated in our work on longitudinal COVID-19 progression [128], can be modeled as a sequence and provide valuable insights into disease progression.

Modeling sequential data presents challenges due to variations in sequence length and the need to capture both short-term and long-term dependencies. Deep learning approaches for time series analysis primarily include Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformers. RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), maintain hidden states that propagate information from previous time steps [207]. Convolutional architectures, like Temporal Convolutional Networks (TCNs) [138], leverage hierarchical convolutions to capture temporal patterns. More recently, Transformer-based methods have emerged as a powerful alternative for modeling sequential data points [245]. Some signals can also be converted to images and then processed by CNNs. This has been applied to audio signals from surgery by modeling them as spectrogram images [210]. Hybrid approaches combine the strengths of different architectures by first encoding complex data modalities before feeding the encodings into a sequential model. For instance, Czempiel et al. [56] employed a CNN to encode individual frames of endoscopic videos and then passed the resulting embeddings to a TCN for further processing [56].

3.2.4 Tabular Data

Tabular data is one of the most prevalent forms of data across various domains, including finance, manufacturing, climate science, and healthcare [214]. In medicine, tabular data encompasses a wide range of information, such as categorical variables like encoded diseases and gender, continuous variables like body temperature, and ordinal variables like cancer staging. However, applying deep neural networks to tabular data poses several challenges due to the lack of inherent spatial or temporal relationships between features, the high prevalence of missing values, sparsity, and the heterogeneity of feature types [214].

Traditional machine learning techniques that do not rely on deep learning, such as Random Forests [26] and gradient-boosted decision trees (GBDT) like XGBoost [48], have demonstrated remarkable performance on tabular data. These methods often outperform deep learning approaches in various benchmarks while requiring significantly less training time and computational resources [155].

Despite the aforementioned challenges, researchers have proposed several deep learning architectures for tabular data, including Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and Transformer-based approaches like FT-Transformer [84]. Additionally, methods that leverage self-supervised pretraining, such as SAINT [217] and SCARF [9], have been introduced, claiming to achieve performance comparable to GBDT on certain datasets.

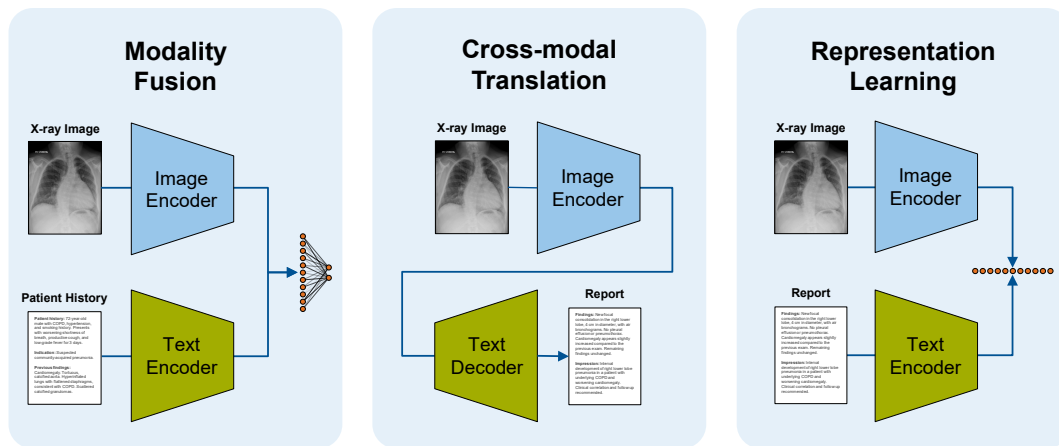


Fig. 3.2. There are three primary types of multimodal deep learning: First, the fusion of modalities (Section 3.3.1) to solve a downstream task, such as classifying findings in a chest X-ray by using both the image and the clinical history written as text. Second, the translation (Section 3.3.2) from one modality to another, like generating a report for a given chest X-ray image. Third, learning multimodal representations (Section 3.3.3) without explicit supervision, such as using contrastive learning.

The intrinsic difficulties associated with deep learning methods for tabular data also pose challenges for fusion techniques based on deep learning. Consequently, alternative approaches have been explored, such as extracting relevant image features as tabular data and processing them alongside other patient data using classical machine learning methods, which will further be discussed in Section 3.3.1.

In summary, while substantial efforts have been made to adapt neural networks for tabular data, classical machine learning methods like GBDT continue to excel in this domain, particularly given the challenges inherent in medical data, such as incompleteness, long-tail distributions, and limited dataset sizes [155]. Nevertheless, deep learning-based models remain highly relevant for multimodal approaches to enable joint optimization and modality interaction, as discussed in the following sections.

3.3 Multimodal Representation Learning

In the previous sections, we explored the most relevant data modalities for clinical decision-making and how deep learning models can be independently trained on these modalities. This section will provide an overview of how multiple modalities can be integrated into deep learning approaches to either improve upon tasks similar to those described for unimodal methods, such as classifying diseases in a patient or to tackle uniquely multimodal tasks like generating radiology reports conditioned by images. Multimodal large language models like our work RaDialog [186], a conversational assistant for radiology report generation, combine many of the aspects of this chapter and will be further discussed in the outlook of this thesis (Chapter 13).

Multimodal deep learning can be broadly categorized into three main paradigms, each addressing different aspects of integrating and leveraging information from multiple data

sources. The first paradigm, modality fusion (Section 3.3.1), combines representations from different modalities to enhance performance on downstream tasks. For instance, the accuracy of classifying findings in a chest X-ray can be improved by considering the image and the patient’s clinical history in text form. The second paradigm, cross-modal translation (Section 3.3.2), aims to convert information from one modality to another, enabling tasks such as generating a radiology report for a given chest X-ray image. The third paradigm, multimodal representation learning (Section 3.3.3), focuses on learning joint representations from multiple modalities during a pre-training phase, often without explicit supervision. This is typically achieved through techniques like contrastive learning, which encourages the model to learn meaningful connections between different modalities. This multimodal representation can then be used for cross-modal retrieval, such as finding similar reports for a given report, or can serve as an initialization for the supervised training of downstream tasks.

3.3.1 Modality Fusion

As discussed in Section 2.1.2 and Section 3.1, integrating heterogenous data sources is essential in the medical decision process and, consequently, in decision support systems based on deep learning. Following Huang et al. [105], modality fusion strategies can be categorized into three main categories: early fusion, joint fusion, and late fusion, as illustrated in Figure 3.3.

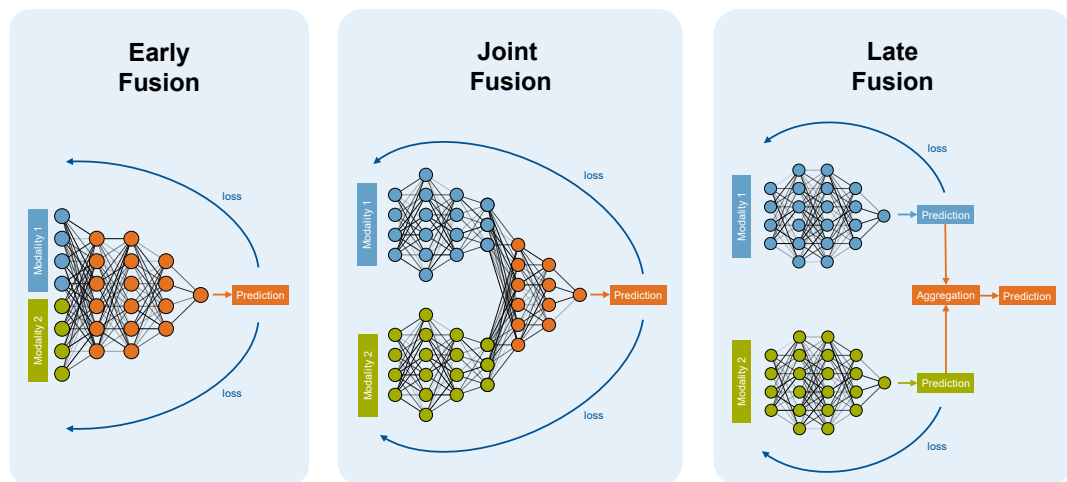


Fig. 3.3. There are three main ways to fuse different modalities of data: early fusion, joint fusion, and late fusion.

Early fusion, or data-level fusion, refers to combining two or more input features of the same dimension into a joint representation, which is then fed to a machine-learning model. The input features can consist of data-level features or features extracted by another model, referred to as feature-level fusion. While logit, probability, or category-level predictions could also be interpreted as feature vectors, we consider this approach late fusion. For effective early fusion, the input features should be on a similar level of abstraction. For instance, to perform early fusion on radiological images and tabular clinical data, the images can be converted to tabular data by extracting radiomics or encoding the presence of radiological findings in a binary vector, as has successfully been demonstrated for COVID-19 by Burian et al. [29]. Another approach for early fusion of multimodal patient data is describing both tabular patient data and extracted radiological findings with text and then processing this text along with

other written patient information like reports and patient history in a language model similar to RaDialog [186] and LENS [20]. Although straightforward, early fusion may not capture the interaction between modalities. In the case of feature-level fusion, the loss of the predictive model is not backpropagated to the feature extraction stage. [105, 197]

Late fusion, also known as decision-level fusion, involves aggregating predictions from individual models trained separately on each modality. Aggregation can be performed using techniques such as averaging, majority voting, or machine learning models like logistic regression, naïve Bayes, or multilayer perceptrons (MLPs). The main advantage of late fusion is that it is flexible, and existing models can be easily combined. However, this approach might not capture low-level interactions between modalities. [105, 146]

Joint fusion, also referred to as deep fusion or intermediate fusion, involves fusing latent representations of neural networks from different modality encoders and processing them together. The key difference between joint fusion and early fusion is that the latent representations on which the fusion happens can be optimized for. Depending on the task and modalities, this fusion can occur at various stages of the model. In Chapter 6, we propose a novel joint feature fusion method for predicting clinical outcomes of COVID-19 patients by leveraging both imaging and clinical data. While joint fusion introduces added complexity, its main benefit is that the feature extraction can be end-to-end optimized for the feature fusion and the downstream task at hand. This allows the model to learn and exploit cross-modal interactions more effectively. [105, 146]

Within the category of joint fusion, two dominant paradigms have emerged: feature-based fusion and token-based fusion. These approaches differ in how they represent and combine information from multiple modalities to learn a unified multimodal representation.

Feature-based Fusion with Pooling

Feature-based fusion involves encoding each modality into dense latent representations using modality-specific encoders, such as convolutional neural networks for images or recurrent neural networks for text. These latent representations globally capture each modality and are then aggregated into a multimodal representation. If the latent vectors have different dimensionalities, they can be projected to the same dimensionality using a simple linear layer before fusion. The actual fusion step merges two or more latent representations into a single representation. This is commonly accomplished using techniques such as element-wise max-pooling, average-pooling (which is equivalent to additive-pooling), or more sophisticated aggregation methods like weighted pooling, attention gates, or bilinear (gated) models [126]. Feature-based fusion allows for modality-specific processing and can effectively capture interactions between modalities at an intermediate level of abstraction.

Token-based Fusion with Transformers

In contrast, token-based fusion approaches, inspired by the success of Transformer [236] and BERT [61] in natural language processing, treat different modalities as a sequence of tokens and apply transformer-based attention mechanisms like self- or cross-attention to these sequences [255]. Architectures like MMBT [125] represent an image as multiple tokens using patch embeddings, similar to Vision Transformers [65], and add these tokens

to the sequence of text or other modalities encoded as tokens. Similarly, LLaVa [147], a recently proposed multimodal LLM, uses a simple yet surprisingly effective projection of image patch embeddings to the language token domain. In RaDialog [186], we use BLIP-2 to project CNN patch embeddings to the language domain and feed them to an autoregressive Transformer decoder. In all these approaches, the transformer encoder learns to attend to and fuse information from the multimodal token sequence. Instead of pooling the features of each modality to global embeddings like in feature-based fusion, the individual tokens preserve the locality of features that can represent words or individual patches of an image. Token-based fusion enables the model to capture complex relationships and dependencies between modalities at a fine-grained level, leveraging the power of self-attention mechanisms. Recently, this approach has been adapted to integrate multimodal medical data into the token sequence of large language models [18].

3.3.2 Cross-Modal Translation

Cross-modal translation involves converting data from one or more modalities to another while preserving task-relevant semantic content and modality-invariant information. In radiology and the clinical domain, this has been applied to report generation, image synthesis, and visual question answering (VQA) [94].

Report generation focuses on producing a clinically accurate and meaningful report based on a given medical image. However, assessing the clinical correctness of generated reports can be challenging, as the text is often unstructured, which is further discussed in Chapter 7.

Cross-modal image synthesis involves generating a corresponding image in a different modality. For instance, a CT image can be generated from an MRI to facilitate registration with other CT images, or contrast enhancement can be applied to stained histology images. In our work U-PET [131], we generate a corresponding FDG-PET image from a given MRI to improve Alzheimer’s disease detection and interpretability. Generating images from a radiology report can be valuable for data augmentation, particularly for underrepresented data distributions. Chambon et al. [39] investigated this approach by conditioning a latent diffusion model with radiology reports to synthesize high-fidelity, diverse X-ray images.

Visual question answering (VQA) involves answering questions about a given image through classification or open-ended responses. Effective VQA relies on the fusion of text (the question) and image information, as discussed in the previous section on multimodal fusion (Section 3.3.1). In our work on Rad-ReStruct [185], we propose modeling the structured reporting task as a sequence of questions following a hierarchy of granularity in the report aligned with the clinical workflow (see Chapter 7).

Current state-of-the-art cross-modal translation models for VQA and report generation, such as RaDialog [186], employ modality fusion techniques, as discussed in Section 3.3.1, followed by text generation using an autoregressive transformer decoder, as described in Section 3.2.1. For image synthesis tasks, diffusion-based denoising models have emerged as the predominant approach, as exemplified by Chambon et al. [39]

3.3.3 Self-supervised Representation Learning

Self-supervised learning is a promising approach to overcome the limitations of supervised methods, which require labeled data for each task. In the medical domain, where high-quality annotations are scarce and expensive, self-supervised techniques can leverage unlabeled multimodal data that is more readily available at large scale [25, 49, 224]. The supervision signal for self-supervised learning comes from the inherent structure and relationships within the data itself. For example, in a hospital database, different data modalities, such as medical images, clinical notes, and laboratory results, are associated with specific patients. A self-supervised system can learn meaningful representations by trying to find similarities and matching pairs across modalities. This provides a strong supervision signal inherent in the data, as their semantic consistency can be exploited. Besides cross-modal retrieval with similarity metrics, it can be used as pre-training, where it serves as an initialization to fine-tune on downstream tasks. These representations can also be used for zero-shot approaches like Xplainer presented in Chapter 9 without further training. Unlike fusion and translational models that optimize for specific downstream tasks, self-supervised models aim to learn task-agnostic multimodal representations broadly applicable to various problems. Inspired by the success of large-scale pre-training in natural language processing with BERT [61], vision-language with CLIP [194], and language generation with GPT-3 [27], there remains significant untapped potential for multimodal foundation models in medicine.

Most self-supervised methods can broadly be categorized into two main pretext tasks: generative and discriminative. Generative approaches obtain a supervision signal by reconstructing or predicting altered input data, such as masked or noise-corrupted samples. Discriminative pretext tasks, on the other hand, involve matching instances in contrastive models like CLIP or employing self-supervised knowledge distillation methods that enforce consistency of representations across augmented views, as used in DINO [35] for images and data2vec [8] for multimodal data. Recent efforts, such as BLIP-2 [142], have combined contrastive and generative cross-modal supervision to adapt vision encoders for image understanding in LLMs, bridging the gap between foundation models trained on separate domains.

Generative and Predictive Self-supervision

Lu et al. [151] extended BERT-style pre-training to vision-language tasks in ViLBERT using separate transformer encoders, while subsequent approaches employed a unified multimodal transformer encoder [241]. This approach has been applied to radiology in MMBERT [124], where masked X-ray images and reports are used to predict the masked tokens of the radiology report. Similarly, Chen et al. [49] adapted the self-supervised BERT pre-training of Vision Transformers in the Masked Autoencoder (MAE) framework to multimodal radiology pre-training by predicting both image patches and masked text tokens, showing a significant increase in performance in VQA and retrieval tasks. Another self-supervision signal can be reconstructing corrupted input images, such as removing added noise, as in our work on counterfactual explanations using diffusion autoencoders [120]. This self-supervision has also been applied to multimodal data in CoDi [227]. Notably, self-supervised generative models can be useful without further fine-tuning.

Learning Image Concepts with Natural Language Supervision

Contrastive language-image pre-training (CLIP) was initially introduced by Zhang et al. [271] for radiology images and was later adapted to large-scale training by Radford et al. [194]. CLIP employs an InfoNCE loss to align the representations of paired images and text descriptions. It has emerged as a powerful technique for generating multimodal text-image representations that have demonstrated significant utility in various downstream tasks, such as classification in a zero- and few-shot setting, captioning, and cross-modal retrieval. While CLIP by Radford et al. [194] was initially trained on general data sourced from the internet, recent efforts have focused on further adapting and applying this approach to enhancing retrieval-based radiology report generation [70]. Building upon these foundations, Boecking et al. [25] further refined the pre-training process by incorporating semantic concepts and discourse characteristics specific to the radiology domain. Similarly, Wang et al. [244] extended the CLIP framework by introducing pre-training on unpaired datasets and incorporating a semantic matching loss. Similar to our work in Chapter 8, Wu et al. [250] introduces a fully supervised contrastive vision-language framework using triplets describing clinical findings in radiology reports.

The success of CLIP and its application in radiology has inspired contrastive pre-training with other modalities, such as retinal fundus images paired with genetic information [223], cardiac MR images paired with tabular data [87], kidney MR images paired with sparse tabular data described with text [13] and electrocardiogram paired with reports [141]. In BioCLIP [267] a large-scale foundational CLIP model was trained on 15 million figure-caption pairs extracted from biomedical literature.



- Most medical data can be represented as text, image, sequential, and tabular data.
- The integration of multimodal medical data is challenging due to heterogeneity, varying dimensionality, missing values, and the combination of dense with sparse data.
- Multimodal data can be fused, translated, or used for self-supervised learning of multimodal patient representations.
- There are fusion strategies on different levels of abstraction: early, joint, and late fusion.

Part II

Modelling Formal and Experiential
Knowledge

Explicit Integration of Exemplar Knowledge in Deep Learning

In Section 2.1.4, we discussed the two primary forms of knowledge that influenced clinical decision-making: formal and experiential. Formal knowledge encompasses the explicit information acquired through medical education, scientific literature, and clinical guidelines, primarily used in analytical decision-making. In contrast, experiential knowledge is gained through hands-on experience in diagnosing and treating patients and is predominantly employed in intuitive decision-making. As we have discovered in Section 2.2, many similarities exist between the cognitive modeling of a clinician's mind and the approaches explored for decision-support systems. Beyond the implicit incorporation of experiential knowledge through training datasets used for optimization, various deep learning methods have been investigated to explicitly utilize this data during inference, similar to a physician recalling relevant past patient cases. Unsurprisingly, the two types of experiential knowledge observed in human cognition, exemplar and prototypical (see Section 2.1.3), have also been explored in the context of deep learning.

This part will first give an overview of the explicit modeling of exemplar knowledge in deep learning and then present two of our methods incorporating this strategy.

Chapter 5 introduces a decision-support system for intoxication prediction based on the limited information a caller provides to a poison control center (PCC) during an emergency. In this system, exemplar knowledge about previous patients in the PCC database is modeled as a population graph and processed by a graph attention network. Furthermore, our model incorporates formal knowledge about symptoms typically associated with the most common intoxications based on etiology, combining this with predictions drawn from exemplar knowledge.

In Chapter 6, we present a method for predicting the outcomes of COVID-19 patients using a population graph to explicitly model the experiential knowledge about previously treated patients. We employ multimodal information about the patients to model their relevance within the graph for new patients at inference, considering all clinically available modalities such as CT images, clinical data, and radiomics for outcome prediction.

4.1 Prototypical Networks

The prototypical knowledge model in cognitive psychology (Section 2.1.3) is reminiscent of the k -means clustering technique in traditional machine learning. A similar concept has recently been incorporated into deep learning, with the ProtoPNet model by Chen et al. [45] drawing inspiration from how clinicians explain their reasoning process. Their explanatory approach follows the "this looks like that" intuition, suggesting that experts identify parts of an

image that resemble typical patterns associated with a particular finding. This aligns with the cognitive model of prototype-based experiential knowledge retrieval in intuitive reasoning.

Since the introduction of ProtoPNet, researchers have explored similar methods for various medical applications, such as aortic stenosis classification [235], X-ray classification [127], and diabetic retinopathy detection [96]. Unlike post-hoc interpretation methods, which will be discussed in more detail in Chapter 10, these prototype-based approaches aim to make the intrinsic reasoning process transparent by assigning semantic concepts to the identified prototypes. Recently, Wolf et al. [247] introduced Shapley values of the similarity metrics to ProtoPNet, to improve the faithfulness of the provided visual explanations.

4.2 Retrieval of Exemplar Knowledge

Long et al. [149] proposed a retrieval augmented classification (RAC) approach that models the exemplar retrieval process in intuitive reasoning (Section 2.1.3). This method incorporates an explicit retrieval module that selects relevant training samples during inference, leading to improved performance on long-tail data distributions. Drawing inspiration from the success of retrieval augmented generation (RAG) in natural language processing [140], Blattmann et al. [22] introduced the concept of sample retrieval for enhancing image generation using diffusion models, resulting in higher quality outputs with reduced memory and computational requirements. In the context of cross-modal retrieval, Endo et al. [69] explored retrieving the most similar report for a given chest X-ray. Their findings suggest that generating a report by sampling the most similar report sentences from multiple similar patients yields competitive results in report generation.

4.3 Modelling Exemplar Knowledge with Graphs

An alternative approach to modeling exemplar knowledge involves representing data samples in a graph structure. Instead of performing a simple similarity-based retrieval of k nearest samples, the retrieval and processing tasks are carried out on the graph. Graph Convolutional Networks (GCNs) have shown great promise in medical applications, especially in optimizing medical image information processing.

Parisot et al. [179] were the first to employ GCNs on population graphs for enhancing the prediction of autism and Alzheimer's disease. They also found that the type of patient information used for graph construction significantly impacted results [177]. Subsequent research aimed to reduce the dependence on graph construction methods, with Anirudh and Thiagarajan [6] proposing a bootstrapping method and ensemble learning for GCNs. Cosmo et al. [54] introduced a graph learning method that integrated both tabular and imaging information to optimize GCN training. Besides population graphs, GCNs have also been applied to medical image segmentation tasks [159, 216, 229, 249]. While the studies mentioned above mainly used pre-extracted image features, Burwinkel et al. [30] proposed a method that directly applied GCNs to image data. They demonstrated that end-to-end optimization within a GCN can improve performance due to optimized feature extraction from

images and clinical patient information. Moreover, their approach enabled more effective modeling of inter-class relationships within the graph. This concept will be further expanded, and its implications will be discussed in detail in Section 6.3.

Intoxication Prediction with Population Graphs and Medical Knowledge

Contents

5.1	Introduction	33
5.2	Methodology	34
5.2.1	Population Graph Processing	35
5.2.2	Integration of Formal Knowledge	36
5.3	Experimental setup	37
5.4	Results and Discussion	38
5.4.1	Ablative Testing and Baselines Comparison	38
5.4.2	Comparison with Clinicians	39
5.5	Conclusion	40

5.1 Introduction

Intoxication is a significant global health concern, causing millions of deaths and disability-adjusted life-years (DALYs) annually. In 2016, alcohol abuse resulted in 2.8 million deaths and 99.2 million DALYs, accounting for 4.2% of all DALYs. Other intoxications contributed an additional 31.8 million DALYs and 451,800 deaths worldwide [59]. Rapid detection of the underlying toxin and appropriate treatment are critical to prevent severe damage to organs or fatalities in cases of intoxication [134]. Poison control centers (PCCs), like the PCC at the toxicology department of the university hospital Klinikum rechts der Isar in Munich, were established to assist medical professionals and the public in classifying and managing intoxications. When the substance causing the patient’s condition is unknown, the medical doctor (MD) at the emergency hotline of the PCC must diagnose the patient based solely on reported symptoms, without direct patient contact. This task is challenging for inexperienced MDs due to variations in symptom descriptions, individual patient responses, and confounding factors from comorbidities. Current clinical decision support systems (CDSS) in toxicology, primarily rule-based expert systems [14, 57, 150], are sensitive to input variations and do not consider demographics like age, gender, weight, or area of living, which are crucial for accurate diagnosis, as discussed in Section 2.2.

To this end, we propose a novel method [31, 266] that leverages Graph Convolutional Networks (GCNs) [58, 129] to incorporate clinical and demographic information into the diagnostic process. Analogous to the exemplar memory in clinical reasoning (Section 2.1.3),

our graph-based approach explicitly models patients and their similarities in a patient population graph, where each patient is represented as a node, and patients are connected based on how similar their tabular metadata is [178]. GCNs perform local filtering of data structured in a graph, in a similar way as Convolutional Neural Networks (CNNs) process grid data (see Section 3.2.2). They have demonstrated success in various medical applications, as discussed in Section 4.3. We base our model on Graph Attention Networks (GATs) [237], which employ an attention mechanism for feature aggregation compensating for imperfect graph neighborhoods.

In addition to considering previously diagnosed patients, our approach integrates formal knowledge (see Section 2.1.4) about toxins and their symptom manifestations from medical textbooks. For this, we employ a literature-matching network that learns a mapping of reported patient symptoms to typical symptoms described in medical literature. This combination of experiential knowledge, represented by the patient population graph, and formal knowledge, incorporated through the literature-matching network, aims to model the complementary way of thinking employed by experienced clinicians in the diagnostic process as described in Section 2.1.3.



Contributions:

- We propose ToxNet, a novel graph-based architecture for improved toxin prediction integrating patient symptoms, demographics, and formal knowledge.
- We develop a conceptual mapping of ambiguous symptoms reported to well-defined textbook symptoms, incorporating medical literature knowledge into the model.
- We evaluate ToxNet on a large PCC dataset and compare its performance against medical experts on a real-life test set.

5.2 Methodology

ToxNet performs toxin classification using patient symptom vectors \mathbf{P} , non-symptom metadata \mathbf{Q} , and literature symptom vectors \mathbf{H} in a graph-based approach. The objective function $f(\mathbf{P}, G(\mathbf{P}, \mathbf{Q}, \mathbf{E}), \mathbf{H}) : \mathbf{P} \rightarrow \mathbf{Y}$ is optimized, where $G(\mathbf{P}, \mathbf{Q}, \mathbf{E})$ is a graph with vertices containing symptoms \mathbf{P} and metadata \mathbf{Q} , edges \mathbf{E} represent connections between vertices, and \mathbf{Y} is the set of toxin classes. Each patient has a binary symptom vector \vec{p}_i . Each toxin has a literature symptom vector \vec{h}_i , forming the sets: $\mathbf{P} = \vec{p}_1, \vec{p}_2, \dots, \vec{p}_M, \vec{p}_i \in 0, 1^{F_P}$, $\mathbf{H} = \vec{h}_1, \vec{h}_2, \dots, \vec{h}_C, \vec{h}_i \in 0, 1^{F_H}$, where M is the number of patients, C is the number of toxin classes, and F_P and F_H are the dimensions of the patient and literature symptom vectors, respectively. The patient metadata is contained in $\mathbf{Q} = \vec{q}_1, \vec{q}_2, \dots, \vec{q}_M$. For each graph vertex, the patient symptom vector \vec{p}_i is concatenated with the corresponding metadata \vec{q}_i to form \mathbf{X} with vectors \vec{x}_i of dimension F . Edges \mathbf{E} are constructed based on the similarity of metadata between nodes.

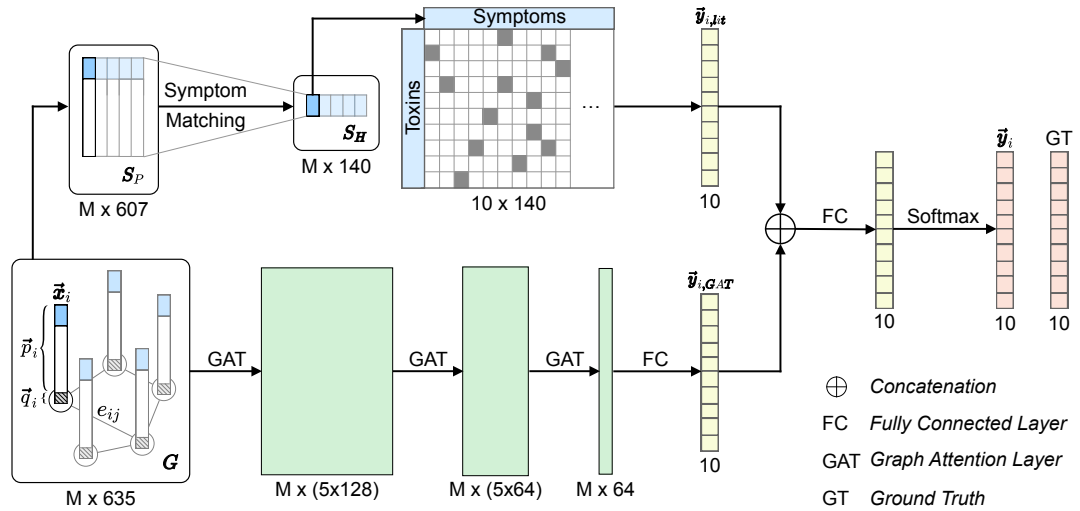


Fig. 5.1. The patient symptom vectors are processed in parallel by a population Graph Convolutional Network (GCN) and a symptom-matching network. The population GCN, based on Graph Attention Network (GAT) layers, considers both reported symptoms and patient demographics to refine patient representations by considering features of similar patients. Simultaneously, the symptom-matching branch maps patient-reported symptoms to textbook descriptions, integrating formal knowledge into the model. The resulting outputs from both branches are then combined to classify the toxin. [31] *Reproduced with permission from Springer Nature.*

The model processes reported symptoms through three graph attentional layers and a parallel literature-matching module. Finally, the outputs from both branches are fused, and using this joint representation the toxin is classified.

Symptom Vectors

Each symptom vector is a binary encoding of the presence or absence of all considered symptoms. The dimensions F_P and F_H represent the total number of unique symptoms S_P and S_H found in all cases in the PCC database and the intoxication literature, respectively. Since reported symptoms may not mentioned in the literature, $F_H < F_P$ and $S_H \subseteq S_P$. In each patient symptom vector \vec{p}_i the first F_H entries correspond to the symptoms in S_H , ensuring that the literature symptoms are consistently represented across all patient vectors.

5.2.1 Population Graph Processing

Graph Construction

The edges E define the neighborhood of each vertex \vec{x}_i , which is formed by concatenating the patient symptom vector \vec{p}_i and the corresponding metadata \vec{q}_i . The neighborhood N_i of \vec{x}_i consists of all vertices \vec{x}_j connected to \vec{x}_i by an edge $e_{ij} \in E$. These neighboring vertices $\vec{x}_j \in N_i$ are aggregated refining the representation of \vec{x}_i within each graph attentional layer. An edge e_{ij} is established between vertices \vec{x}_i and \vec{x}_j when their respective metadata is consistent, ensuring that the graph captures meaningful connections between patients with similar characteristics.

Graph Processing

Following Veličković et al. [237], the GAT layer updates the representation of each vertex \vec{x}_i in \mathbf{X} by applying a shared learnable linear transformation $\mathbf{W} \in \mathbb{R}^{F' \times F}$, resulting in a new representation with dimension F' . For each neighbor $\vec{x}_j \in N_i$, an attention coefficient α is computed using a shared attention mechanism a , which represents the importance of \vec{x}_j in updating \vec{x}_i . The attention coefficient is calculated as $a(\mathbf{W}\vec{x}_i, \mathbf{W}\vec{x}_j) = \vec{a}^T[\mathbf{W}\vec{x}_i \parallel \mathbf{W}\vec{x}_j]$, where \parallel denotes concatenation and $\vec{a} \in \mathbb{R}^{2F'}$ represents a single feed-forward layer. To normalize the attention coefficients and facilitate comparability, a leakyReLU activation σ is applied, followed by a softmax function over all coefficients corresponding to the neighbors in N_i for each \vec{x}_i .

$$\alpha_{ij} = \frac{\exp(\sigma(\vec{a}^T([\mathbf{W}\vec{x}_i \parallel \mathbf{W}\vec{x}_j])))}{\sum_{r \in N_i} \exp(\sigma(\vec{a}^T([\mathbf{W}\vec{x}_i \parallel \mathbf{W}\vec{x}_r])))} \quad (5.1)$$

To update \vec{x}_i , the transformed feature representations $\mathbf{W}\vec{x}_j$ of its neighbors are weighted by their corresponding attention coefficients α_{ij} and aggregated through summation to obtain the new representation \vec{x}'_i . The graph attentional layer repeats this process K times using independently learned transformations \mathbf{W}^k , referred to as heads, to stabilize the predictions and capture different attention patterns via head-specific attention coefficients α^k . The resulting representations \vec{x}'_i from each head are concatenated (denoted by \parallel) to form the final updated representation:

$$\vec{x}'_i = \parallel_{k=1}^K \sigma \left(\sum_{j \in N_i} \alpha_{ij}^k \mathbf{W}^k \vec{x}_j \right) \quad (5.2)$$

In this equation, K represents the number of attention heads employed, and α_{ij}^k denotes the attention coefficient computed by head k for vertices \vec{x}_i and \vec{x}_j [237]. Using multiple attention heads allows the graph attentional layer to capture diverse aspects of the neighborhood structure and improve the stability of the learned representations.

5.2.2 Integration of Formal Knowledge

Literature Symptom Matching

For each toxin class c_i in the set of all toxins C , the literature provides a list of frequently observed symptoms, which are encoded in a binary symptom vector \vec{h}_i for every toxin. To map the patient symptom vectors \mathbf{P} to the literature symptoms, we introduce a dedicated symptom matching layer $\mathbf{W}_{\text{symp}} \in \mathbb{R}^{F_H \times F_P}$. This layer learns an interpretable transfer function that offers insights into symptom correlations and explicitly incorporates formal knowledge from the literature. Given the structure of the symptom vectors, the first F_H entries of each patient symptom vector \vec{p}_i correspond to the literature symptoms. To preserve these entries during the matching process, we initialize the first F_H learnable parameters of \mathbf{W}_{symp} with the identity matrix I_{F_H} and freeze the diagonal elements during training. This ensures that each symptom s in S_H is mapped to itself. The remaining symptoms, which are only reported by patients but not found in the literature, are transformed into a representation with a dimension consistent with the symptoms in the literature. As a second transformation, we define a literature layer $\mathbf{W}_{\text{lit}} \in \mathbb{R}^{C \times F_H}$, where the i -th row is initialized with \vec{h}_i for all classes C and remains fixed during training. The resulting transformation $\vec{y}_{i,\text{lit}} = \mathbf{W}_{\text{lit}} \cdot \sigma(\mathbf{W}_{\text{symp}} \vec{p}_i)$ maps the patient symptoms onto the toxin classes by explicitly leveraging the literature information.

Branch fusion

The output of the final GAT layer is passed through a fully connected (FC) layer to obtain \vec{y}_i, GAT , a representation with dimension C corresponding to the number of toxin classes. The GAT representation \vec{y}_i, GAT and the literature representation $\vec{y}_{i, lit}$ are then concatenated, activated, and passed to a final linear layer for classification. \vec{y}_i represents the final softmax probability of each toxin class for the given patient symptoms and metadata.

5.3 Experimental setup

Poison Control Center Dataset

The dataset, extracted from the PCC database spanning the years 2001-2019, consists of 8,995 patients with confirmed mono-intoxications, where only one known toxin was present. We selected ten toxin groups: ACE inhibitors (n=119), acetaminophen (n=1,376), antidepressants (selective serotonin reuptake inhibitors, n=1,073), benzodiazepines (n=577), beta-blockers (n=288), calcium channel antagonists (n=75), cocaine (n=256), ethanol (n=2,890), NSAIDs (excluding acetaminophen, n=1,462), and opiates (n=879). In addition to having similar actionable treatment implications, these toxin groups were chosen based on their frequency, clinical distinctiveness, severity, and the importance of accurate identification. The classes are imbalanced, reflecting the varying prevalence of different intoxications in real-world scenarios (e.g., the high frequency of alcohol intoxication). In addition to patient symptoms, metadata such as age group (child, adult, elder), gender, etiology, point of entry, weekday, and year of intoxication is available for each case. These meta-features were selected to construct the patient population graph based on their contribution to the best performance.

Population Graph Construction

The patient population graph is constructed based on the selected metadata for each patient. An edge e_{ij} is established between patients \vec{x}_i and \vec{x}_j when their metadata is consistent across the medically relevant parameters mentioned above. This approach results in a sparse graph with meaningful edges, increasing the likelihood of connecting patients with the same type of poisoning. To ensure proper evaluation and prevent information leakage, samples in the training set are only connected to other training samples. In contrast, validation and test samples are connected to training samples through directed edges. This setup allows the validation and test samples to be considered only during their respective phases and not during training. Consequently, the inductive GAT network can perform inference on new, unseen patients while leveraging the graph structure provided by the training set during inference.

Implementation Details

We chose the Adam optimizer with a learning rate of 1e-3 and a weight decay of 5e-4 for training the network and cross-entropy loss. No dropout regularization was applied, and the Exponential Linear Unit (ELU) was utilized as the activation function throughout the network. Each graph attentional layer in ToxNet is equipped with 5 attention heads.

Tab. 5.1. Evaluation of different methods for toxin prediction. The methods are detailed in Section 5.3 (p-value: <0.01 *, <0.005 **). [31] *Reproduced with permission from Springer Nature.*

Method	F1 micro	F1 macro	p micro	p macro
Naive Matching	0.20 ± 0.01	0.13 ± 0.01	**	**
Decision Tree	0.25 ± 0.02	0.23 ± 0.02	**	**
LitMatch	0.47 ± 0.01	0.34 ± 0.02	**	**
MLP with meta	0.54 ± 0.02	0.43 ± 0.02	**	**
GAT	0.63 ± 0.01	0.46 ± 0.02	**	**
ToxNet(S)	0.64 ± 0.01	0.48 ± 0.02	**	**
ToxNet	0.66 ± 0.01	0.53 ± 0.04	-	-

Model Evaluation

We evaluate ToxNet against various benchmark approaches and conduct an ablation study to assess the contributions of different network components. We disable specific network components in the ablation study to evaluate their impact. 'GAT' refers to using only the GAT pipeline of ToxNet, 'LitMatch' refers to using only the literature-matching branch, and 'MLP with meta' represents a standard MLP. Both 'GAT' and 'MLP with meta' receive symptom vectors and metadata as input. At the same time 'LitMatch' uses symptom vectors and explicitly encodes the literature vectors (Section 5.2). We also test a sequential setting (ToxNet(S)) where literature matching is performed first, and the learned features are then passed to the GAT. All experiments use 10-fold cross-validation, each containing 10% of the data as the test set. The remaining 90% is further divided into 80% for training and 20% for validation. In addition, we compare ToxNet with 10 MDs on the same unseen subset of the full test data. This subset consists of 25 individual cases for each MD and an additional 25 identical cases for all MDs, totaling 275 cases (250 + 25). This setup allows for statistical analysis of a larger case set and an assessment of inter-variability among MDs to differentiate between easy and challenging cases.

5.4 Results and Discussion

5.4.1 Ablative Testing and Baselines Comparison

Table 5.1 presents a comparison of the F1 micro and macro scores achieved by various benchmark approaches and our proposed method, ToxNet, for the task of toxin classification. As a lower baseline, the Naive Matching approach simply selects the toxin with the highest overlap between the symptoms described in the literature and the patient's symptoms. Additionally, a decision tree model was trained using the literature symptoms and then applied to the patient's symptoms for classification. The suboptimal performance of both the Naive Matching and decision tree models indicates that relying solely on the available literature is insufficient for accurate toxin classification. In contrast, the LitMatch neural network branch of our approach,

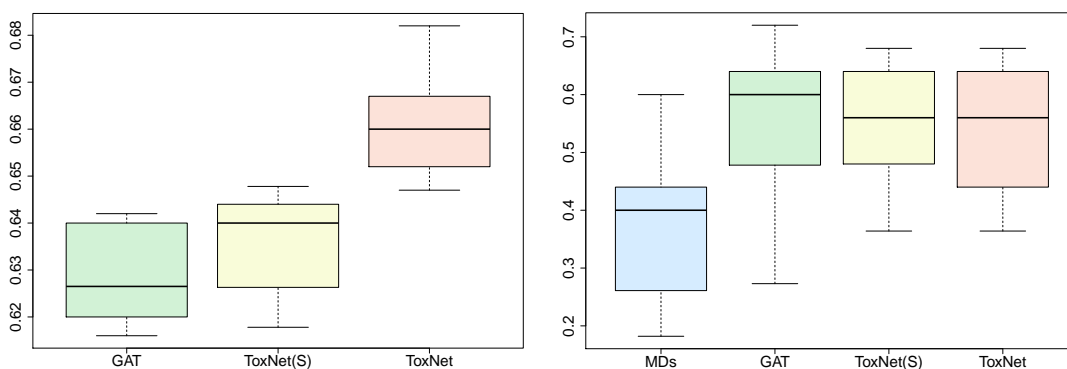


Fig. 5.2. **Left:** Evaluation of ToxNet and various baseline approaches using 10-fold cross-validation. **Right:** Assessment of ToxNet and baseline methods compared to the performance of MDs across 10 distinct datasets, each assessed by a single MD. [31] *Reproduced with permission from Springer Nature.*

which explicitly incorporates literature knowledge, achieves better results while maintaining the ability to leverage this information.

To provide a fair comparison, we trained a Multi-Layer Perceptron (MLP) with three hidden layers (5×128 , 5×64 , and 64 hidden units, respectively) on the patient data for prediction, matching the architecture of the Graph Attention Network (GAT) used in our approach. The patient’s metadata was concatenated to their symptom vector, ensuring that both the MLP and GAT had access to the same information. Comparing the performance of the MLP to a standard GAT network demonstrates that utilizing metadata within our graph-based method leads to a substantial improvement in classification performance, emphasizing the value added by the graph structure.

Furthermore, the incorporation of literature information through our proposed ToxNet method enhances performance even further, despite the literature data alone being relatively uninformative for the task when used in isolation. This improvement suggests a synergistic effect between the patient data and the literature information, and refining the literature might lead to an even greater performance boost. We conducted separate evaluations of both pipelines within ToxNet (GAT and LitMatch) to identify their individual contributions, as described above. Our experiments revealed that the parallel setting of ToxNet slightly outperforms the sequential setting (ToxNet(S)). These results are visually represented in the boxplot in Figure 5.2 (left).

5.4.2 Comparison with Clinicians

To assess the performance of our method in comparison to medical experts, we conducted a survey involving ten MDs from the toxicology department of the Klinikum rechts der Isar in Munich. Each MD was tasked with classifying 50 intoxication cases that were divided as described previously. Among the participating doctors, six were assistant doctors in the toxicology department, while the remaining four were specialists in pharmacology and toxicology. Figure 5.2 (right) presents a box plot comparing the performance of the 10 MDs with various benchmark methods and our proposed ToxNet method on ten individual sets of 25 cases each, totaling 250 cases. The results demonstrate that all three graph-based

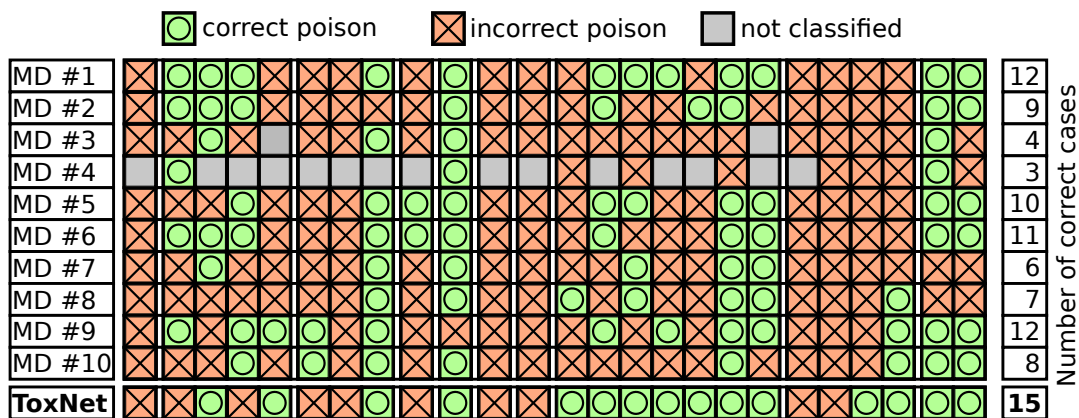


Fig. 5.3. Inter-variability among clinicians and comparison to ToxNet’s performance. The toxin categories are arranged in alphabetical order, with each group separated by a small gap. [31] *Reproduced with permission from Springer Nature.*

approaches outperform the MDs, which can be attributed to the optimized utilization of metadata. Although the performance improvement of ToxNet compared to GAT is not as substantial for this small subset of the full test set, the overall performance exhibits greater stability, as evidenced by the smaller margins.

To further investigate the inter-variability among the MDs, we conducted a detailed study on the 25 cases evaluated by all doctors, as shown in Figure 5.3. Except for one case, our method predicted every toxin correctly, which was also identified by a majority of MDs. Moreover, our method successfully classified eight cases where only half or fewer MDs made correct predictions. These findings highlight the ability of our proposed ToxNet architecture to reliably predict common cases at an expert-level performance while also demonstrating high prediction stability on cases that prove challenging for most doctors. Even when compared to the two best-performing MDs, who correctly classified 12 cases, our method achieved a total of 15 correct toxin predictions. It is worth noting that all doctors and our method misclassified six cases, which can be attributed to inconsistent documentation and incompleteness in the data samples, highlighting the intrinsic challenges associated with real-world medical data.

5.5 Conclusion

In this study, we introduced ToxNet, a novel method to enhance toxin prediction by incorporating patient symptoms and demographics with domain-specific formal knowledge from literature. Our experimental results demonstrate that leveraging metadata within the graph structure of ToxNet leads to significantly higher scores than the compared methods. Through our evaluation, we showed that simply concatenating the metadata to the patient symptom vector for early fusion is insufficient and that the increase in performance can be attributed to the employment of a patient population graph. Furthermore, we introduced a symptom-matching branch that enables the explicit incorporation of textbook knowledge and integrated it, resulting in further improvements to the overall network performance. Integrating experiential knowledge from patient cases with formal knowledge from literature aligns with the combination of exemplar and formal knowledge described in Section 2.1.4.

Although textbook knowledge alone was not informative enough for satisfactory classification, we demonstrated that its parallel integration with our graph network led to synergistic effects and enhanced classification accuracy.

To validate the performance of ToxNet, we conducted an evaluation against ten medical doctors (MDs) with varying levels of experience. Considering the high inter-rater variability among experts, our method exhibited more robust predictions on both common and challenging intoxication cases. These findings underscore the potential of ToxNet as a CDSS in the time-critical and high-stake use case of toxin prediction, where such a CDSS can support the clinician in finding relevant cases from the past and integrating them in the decision-making process. In the next chapter, we will show that the retrieved patients and the attention of the graph network on them can be visualized, making this reasoning more transparent.

COVID-19 Outcome Prediction with Multimodal Population Graphs and Joint Pathology Segmentation

Contents

6.1	Introduction	43
6.2	Related Work	44
6.2.1	Integrating Imaging and Tabular Data	45
6.2.2	Graph Convolutional Networks	46
6.2.3	Multitask Learning	46
6.3	Method	47
6.3.1	Graph-based Image Processing	47
6.3.2	Segmentation, Image Features, and Radiomics	49
6.3.3	Multimodal Feature Fusion	50
6.3.4	Patient Outcome Prediction	50
6.4	Experiments	51
6.4.1	Multimodal COVID-19 Datasets	51
6.4.2	Implementation Details	55
6.4.3	Ablative Testing and Baselines	57
6.4.4	Metrics	58
6.5	Results and Discussion	59
6.5.1	Population Graph Construction	59
6.5.2	U-GAT Evaluation	59
6.5.3	Interpretability and Graph Attention	65
6.5.4	Challenges and Outlook	66
6.6	Conclusion	66

6.1 Introduction

The first wave of the COVID-19 pandemic posed unparalleled difficulties for healthcare infrastructures, with an exponential surge in cases overwhelming intensive care units (ICUs) and presenting scenes unwitnessed in modern medicine [198, 203]. Optimizing hospital resource planning, such as ICU beds, ventilators, and medical staff, becomes critical during such emergencies. Therefore, correctly anticipating treatment necessity and potential outcomes is essential for effective patient management. However, predicting this is challenging when faced with a novel disease, limited understanding, and highly heterogeneous data. This overload on healthcare facilities, paired with the complexity of the data, highlighted the need

for decision support systems that can predict patient outcomes and help triage by making use of all available patient data.

At the onset of the COVID-19 pandemic, numerous parameters were obtained and documented upon a patient's hospital admission, including sex, age, body weight, symptoms, co-morbidities, blood cell counts, inflammatory markers, biochemical values, and cytokine profiles [29]. This tabular data (see Section 3.2.4), along with radiological images such as radiographs or computed tomography (CT) scans, was available within the first hours of a patient's arrival, making them ideal for early triaging and outcome prediction. Given the inherent uncertainty in clinical findings and diagnostic tests, as highlighted in Section 2.1.2, it is crucial to integrate all available patient information from various sources to effectively diagnose and treat patients. Furthermore, as discussed in Section 2.1.3, clinicians often rely on exemplar knowledge, which refers to experience with similar patients who have been previously treated, to diagnose disease outcomes. As discussed in Section 4.3, structuring and retrieving these patients' representations can explicitly be modeled using a patient population graph [177], representing relationships between patients with similar characteristics and outcomes.

Drawing inspiration from these clinical reasoning methods and the need for integrated diagnostic support, we propose a decision support system that performs multimodal data analysis to create a patient population graph, which is then utilized in a graph attention network to refine patient outcome predictions by considering similar patients [121]. The similarity metric, attention mechanism, and generated pathology segmentations offer added insight into the decision-making process by enabling direct observation of the weighting of clinical features and the most influential patients in the prediction process.



Contributions:

- We introduce U-GAT: an end-to-end, graph-based method leveraging multimodal data for patient outcome prediction in COVID-19.
- We employ a multitasking approach with simultaneous segmentation and classification using U-Net and Graph Attention Network (GAT).
- We propose an interpretable, multimodal patient similarity metric for population graph construction and effective batch selection.
- We develop a novel equidistant image sampling method for end-to-end training of volumetric image feature extraction in a graph convolutional setting.
- We evaluate the approach on a private dataset acquired at Klinikum rechts der Isar and an external, publicly available dataset.

6.2 Related Work

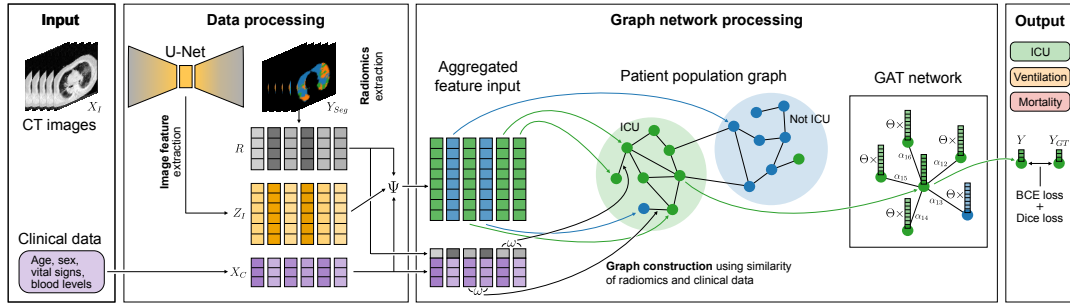


Fig. 6.1. U-GAT is a multitasking model that segments pathologies and uses this information to predict various patient outcomes. It integrates image features (Z_I), radiomic features (R), and clinical metadata (X_C) and is end-to-end optimized to extract the most relevant features for disease prognosis. The clinical metadata includes information such as age, sex, vital signs, and blood levels. The model segments disease-affected areas (Y_{Seg}) in CT images (X_I), which helps in extracting radiomic features (R) and regularizes the extraction of image features (Z_I). These features are fused into a multimodal vector representation using a function Ψ . Test patients are clustered with training patients in a graph based on the similarity between their radiomic and clinical data features defined by ω . A Graph Attention Network (GAT) then refines the features to predict the most probable outcome (Y) using a learned linear transformation Θ and patient attention coefficients α_{ij} . Outcome classification is supervised using binary cross-entropy (BCE) loss, while the Dice loss is used for the auxiliary segmentation task. Applied to COVID-19, U-GAT segments pathologies in the lung CT image and predicts outcomes such as ICU admission, ventilation needs, survival, or severity. [121]

6.2.1 Integrating Imaging and Tabular Data

This section reviews methods used for integrating imaging and non-imaging data to support clinical decision-making, as discussed in the Section 3.3.1, and gives an overview of the different data fusion strategies that have been proposed for COVID-19 in particular.

One early approach to connecting features from multiple modalities was introduced by Perez et al. for visual reasoning tasks [187]. A Feature-wise Linear Modulation (FiLM) layer affinely transformed the output of a Convolutional Neural Network (CNN) with a learned scaling and shifting factor using the text of the input question. Similarly, Dynamic Affine Feature Map Transform (DAFT) [248] combines the features of 3D brain T1-weighted MRI scans and non-imaging biomarkers for Alzheimer’s prediction. DAFT affinely transformed the imaging features extracted by a 3D Fully CNN by a learned scaling and shifting factor using nine non-imaging features, such as age, sex, and genetic factors. A multi-headed cross-attention block has been recently proposed to fuse imaging and tabular data for skin lesion classification using a transformer architecture [32]. Moreover, Duanmu et al. [67] combined breast MRI scans and clinical biomarkers to predict chemotherapy response.

Following the categorization of fusion strategies for multimodal data presented in Section 3.3.1), our approach employs joint fusion. Unlike early and late fusion methods, this allows us to optimize the latent representation of each modality for fusion. Due to the urgent need to integrate multimodal data in CDSSs during the COVID-19 pandemic, several of these strategies have been explored for COVID-19.

Early Fusion

Early fusion of features has been widely applied in methods integrating imaging and non-imaging data for COVID-19 detection and patient outcome prediction [29, 33, 40, 99, 115, 226, 256]. Chassagnon et al.[41] showed the effectiveness of non-imaging and extracted imaging features in an ensemble of machine-learning models for the outcome prediction of COVID-19. This is confirmed by the findings of Shiri et al.[213] on COVID-19 survival prediction by combining lesion radiomics and clinical data. Gong et al. [83] also demonstrated improved prediction of severe COVID-19 outcomes by adding blood test results to other clinical features and extracted radiomics.

Late Fusion

Ning et al.[170] applied late fusion with penalized logistic regression and reported an improvement in both COVID-19 severity and mortality prediction compared to CNN and non-imaging Multilayer Perceptron (MLP) models alone. Tariq et al.[228] investigated different fusion methods to predict hospitalization of COVID-19 patients and found the early fusion of electronic medical record features to be the most effective strategy for this task.

Joint Fusion

To the best of our knowledge, at the time of publishing this work [121], we were the first to propose a joint fusion method combining imaging and non-imaging data to predict ICU admission, ventilation, mortality, or severity for COVID-19.

6.2.2 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) have been adapted for COVID-19 diagnosis, primarily focusing on disease detection. Wang et al.[240] and Yu et al.[265] constructed graphs based on the similarity of extracted CT image features and classified nodes for the detection of infiltrates. Song et al.[218] and Liang et al.[144] incorporated additional features, such as an acquisition site, to improve COVID-19 detection. Saha et al.[204] did not use GCN for population graph processing but instead converted edges detected in chest CT and X-ray images to graphs for COVID-19 detection. Huang et al.[103] used GCNs to refine COVID-19 segmentations. Di et al. [62] employed an uncertainty-vertex hypergraph to classify community-acquired pneumonia and COVID-19. Our work is the first to propose a graph-based patient outcome prediction method by leveraging a population graph combining end-to-end chest CT feature extraction and tabular patient data.

6.2.3 Multitask Learning

Radiological studies [29, 52, 213, 239] on the prognosis of COVID-19 patients have demonstrated a strong link between disease burden and patient outcomes, such as the probability of ICU admission. Different deep learning methods have investigated multitasking approaches [137, 157, 260] to use the correlation between pathological tissue presence and patient health status. However, most of the proposed multitask methods focus on the joint detection of COVID-19 infection and binary segmentation of related pathologies in lung CT

images [4, 5, 10, 77, 251]. They are not concerned with outcome prediction or segmentation of different types of COVID-19 pathologies. In a similar direction, some works [82, 93] applied multitask learning to jointly estimate the severity of COVID-19 and solve related classification and segmentation tasks. Most comparable to our approach, Näppi et al. [168] used extracted bottleneck features of a pretrained U-Net to predict COVID-19 progression and mortality. However, their method differs from our proposed approach in several key aspects. Firstly, they did not optimize the image feature extraction end-to-end, which is a crucial component of our method. Secondly, they did not incorporate clinical patient data, which we believe is essential for accurate patient outcome prediction. Lastly, they did not employ a population graph to model previous patients, a core feature of our method that contributes to its novelty and effectiveness in predicting patient outcomes.

6.3 Method

We propose U-GAT, illustrated in Figure 6.1, that offers a multimodal approach to outcome prediction by forming a holistic patient representation including all relevant data, such as CT images (X_I), radiomics (R) and clinical information (X_C). In the context of COVID-19, we focus on predicting three key outcomes for patients admitted to the hospital: the need for ICU admission, the need for mechanical ventilation, and patient survival (for our in-house dataset). For the iCTCF dataset, we predict COVID-19 severity. As an auxiliary task, we incorporate the segmentation of COVID-19 pathologies for regularization and localized supervision signal. From the segmentation output, we also derive radiomic features (R) that quantify the relative burden of each pathology class on the lungs. To exploit the synergies between image segmentation and outcome prediction tasks, we combine the image understanding capabilities of U-Net with the analytical strengths of Graph Convolutional Networks (GCNs). We use graph processing to not only consider an isolated patient at test time but also use similar patients for feature refinement. This population graph is constructed based on the similarity of clinical patient data (X_C) and radiomic features (R). The model is trained end-to-end, enabling the joint optimization of image feature representation learning, U-Net image segmentation, and graph data processing. Before training, the graph is pre-computed, while at test time, new patients are dynamically attached to the existing graph using the inferred radiomics and patient information.

6.3.1 Graph-based Image Processing

We use spatial graph convolutions to enable inference on unseen data samples without the need for retraining the entire network, like in spectral methods. As discussed in Section 3.3.1, integrating image data (X_I) with other modalities is crucial for comprehensive patient outcome prediction. When using GCNs, image-based features are typically extracted in a separate step, either manually or using a pretrained CNN. These features are then processed within the graph network. Although this approach reduces the memory requirements for imaging data, it limits the potential for end-to-end optimization, a key aspect of our proposed method. Burwinkel et al. [30] demonstrated that the end-to-end image feature extraction can be improved by processing neighboring images using a graph structure and geometric learning. Using our proposed method, we apply this concept to process the provided CT image information. Each

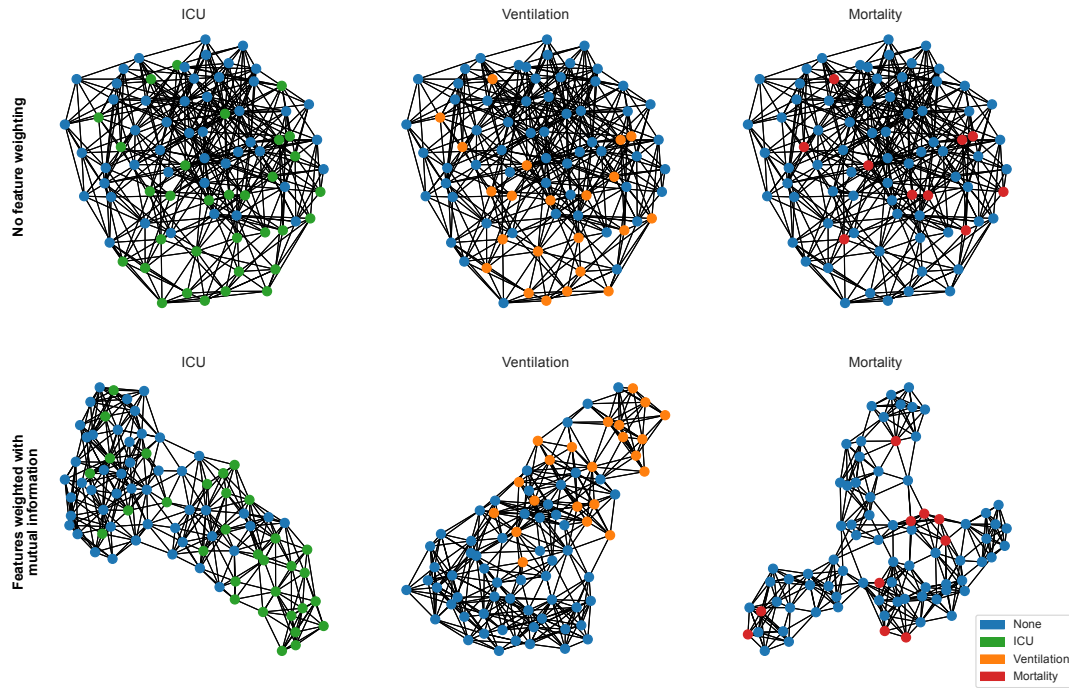


Fig. 6.2. Visualization of the patient population graph for ICU, ventilation, and survival prediction tasks on the KRI dataset. In the top row, each node is connected to its seven nearest neighbors, determined by calculating the Euclidean distance between feature vectors formed by both clinical and radiomic features. We introduce feature weighting in the distance calculation to optimize the graph structure for a specific task without using prior knowledge about the disease. This weighting is based on the mutual information [200] between features and the task at hand, as shown in the bottom row. This gives features relevant to the task a higher priority in forming patient neighborhoods, encouraging aggregation of features from relevant patients. [121]

CT image $x_{I,i}$ is fed into a U-Net architecture to perform segmentation on the individual image slices. The bottleneck features generated by the U-Net (see Section 6.3.2) are then further processed to obtain an image representation $z_{I,i}$, which can be effectively fused with the other modalities and propagated through the graph network.

Equidistant Subsampling of CT Volumes

We propose a simple yet effective approach to address the challenge of high memory demands when utilizing GCNs for end-to-end feature extraction from high-resolution 3D images. This is necessary because GCNs need batch diversity to facilitate feature aggregation within a significant portion of the graph. However, processing the entire CT volume would limit the number of patient instances per batch. Our solution involves equidistant subsampling of S slices per volume along the axial view during the training phase. Given a main axis length of Z , each volume is partitioned into $\lfloor Z/S \rfloor$ stacks of S slices while discarding $(Z \bmod S)/2$ slices on both ends. This approach offers two key benefits: it increases the probability of capturing disease-affected regions and reduces overfitting by distributing limited 3D volume data across multiple patient samples. The entire stack of slices is utilized during testing, covering the complete 3D volume to ensure no pathologies are missing.

Graph Construction Method

We construct a binary, directed graph $G(V, E)$ with a set of vertices V and connecting edges E . Each vertex $v_i \in V$ represents a patient with the information of a stack of CT slices $x_{I,i} \in X_I$ (obtained through the sampling process described in Section 6.3.1), a vector of radiomics features $r_i \in R$ (extracted using the method detailed in Section 6.3.2), and clinical data $x_{C,i} \in X_C$. To build the graph, we first concatenate the clinical data X_C and radiomics features R into a single tabular feature vector. We then calculate the distance ω between pairs of vertices based on these combined features. Finally, each vertex v_i is connected to its k nearest neighbors determined by the calculated distances.

Instead of manually selecting features, we propose a feature weighting approach based on statistical training data analysis. The idea is to assign greater importance to statistically significant features in the distance and, therefore, similarity calculations. Possible weighting schemes include correlation coefficients, such as the Pearson correlation for continuous features, or estimated mutual information [200] between the input features and target labels like Y_{ICU} computed on the training set. The motivation behind using mutual information is to capture both linear and non-linear associations between the features and predicted labels. To mitigate the impact of varying scales across different features, we standardize all features using z-score normalization before calculating the distances between vertices. An alternative approach would be to use another distance metric, like cosine similarity [135], which was not explored in this work. Figure 6.2 illustrates the k-nearest neighbors (KNN) graphs for a training set, both with and without weighting the distance using mutual information.

6.3.2 Segmentation, Image Features, and Radiomics

A segmentation backbone is utilized as both a high-level image feature extractor and a pixel-wise prediction to visually inspect healthy and pathological tissue and calculate radiomics at inference. This backbone can be implemented using any encoder-decoder architecture that generates a compressed bottleneck representation and a segmentation output. In our experiments (see Section 6.4), we employ a modified version of the original 2D U-Net architecture introduced by Ronneberger et al. [199]. The imaging data is processed by sampling S equidistant slices from each input image $x_{I,i}$ (see Section 6.3.1 for slice selection details) and forwarded by the segmentation backbone independently. This results in a 2D segmentation of healthy lung tissue and pathological regions for each slice. The image features for the classification task are obtained by applying global average pooling to the bottleneck features of each slice, reducing the size from $c \times d_1 \times d_2$, where c is the number of channels and d_1 , and d_2 are the spatial dimensions, to a vector of length c per slice. To create a patient-wise representation, the slice-wise image feature vectors are aggregated by computing the element-wise maximum along the stacking dimension. The resulting vector of size c is then passed through a fully connected layer and a leaky ReLU activation function to obtain the latent image representation $z_{I,i} \in Z_I$. Although Goncharov et al. [82] reported improved performance using the U-Net's final feature map instead of the bottleneck, our initial experiments with this approach led to a substantial performance drop. Therefore, we only explored the pooled bottleneck representation further.

Extraction of Radiomic Features

The approach to enrich the clinical data with radiomic features R is inspired by the work of Burian et al. [29]. These radiomics can automatically be extracted from the output of our segmentation network Y_{seg} , e.g., by calculating the disease burden of the lung for a given pathology. Since segmentation masks and radiomics are highly interpretable, this aids the intrinsic interpretability of our method for both patient retrieval in the graph and validation of intermediate image representations.

6.3.3 Multimodal Feature Fusion

Our approach to multimodal representation learning for clinical decision support leverages the image patient data in two complementary ways. First, we extract radiomics features from the segmented pathology regions, converting the image information into tabular data. This allows for an early fusion with patient clinical data, enabling the construction of a patient population graph using both imaging and non-imaging information (see Section 3.3.1). This graph structure captures task-specific patient similarities based on their imaging and clinical characteristics. Second, we perform a joint fusion of latent image features derived from the bottleneck representation of the U-Net architecture with the extracted radiomics features and clinical data. This allows for an end-to-end optimization of the feature fusion most relevant to the task, resulting in comprehensive node features for each patient in the population graph.

The three modalities fused are latent image features $z_{I,i}$ extracted from the input image data $x_{I,i} \in X_I$, radiomics features $r_i \in R$ derived from the segmented regions, and clinical data $x_{C,i} \in X_C$. All of them are propagated through the graph and contribute to the classification task performed for each patient node v_i . Including clinical data X_C is particularly valuable, as it provides complementary information that may not be captured by the imaging-based features alone (see Section 2.1). By incorporating the latent bottleneck features $z_{I,i}$ from the U-Net architecture, our approach enables end-to-end optimization of the image features, allowing for the learning of more expressive representations in addition to hand-crafted radiomics. We apply a learnable linear transformation to the features from each modality to align their dimensionality for equal contribution. This projects them onto a common feature space of dimension F_f . The transformed features are then combined using an aggregation function Ψ to obtain the fused representation $z_{f,i}$ that is subsequently utilized within the graph network:

$$z_{f,i} = \Psi(\sigma(\Theta_I z_{I,i}), \sigma(\Theta_R r_i), \sigma(\Theta_C x_{C,i})) , \quad (6.1)$$

where σ denotes a non-linear activation function, and $\Theta_I \in \mathbb{R}^{F_I \times F_f}$, $\Theta_R \in \mathbb{R}^{F_R \times F_f}$, and $\Theta_C \in \mathbb{R}^{F_C \times F_f}$ are learnable linear transformation matrices that map the input feature dimensions F_I , F_R , and F_C to the common dimension F_f , respectively.

6.3.4 Patient Outcome Prediction

Our proposed method for multimodal representation learning in clinical decision support systems employs graph attention layers (GAT) [237] as the foundation for graph processing. At the same time, our approach allows for other graph convolutional networks. GATs offer several advantages, including effective neighborhood processing, the ability to perform direct

inference on unseen data samples, and the preservation of filter localization while maintaining low computational complexity. Our graph-based approach allows for the effective incorporation of clinical patient data X_C and radiomics R into the learning process by constructing the graph based on the similarity of tabular features and creating a neighborhood $N(i)$ for each fused representation $z_{f,i}$. This enables the aggregation of features from patients with likely similar outcomes across multiple hops in the patient neighborhood. Each Graph Attention Network (GAT) layer in our model aggregates a 1-hop neighborhood within the graph $G(V, E)$ for every vertex v_i . The transformation of a vertex representation within a GAT layer considers the representation itself and the feature vectors of its neighboring vertices. The update of the fused feature representation $z_{f,i}$ within a GAT layer is computed as follows:

$$z'_{f,i} = \parallel_{p=1}^P \sigma \left(\sum_{j \in N(i)} \alpha_{ij}^p \Theta^p z_{f,j} \right) \quad (6.2)$$

where α_{ij}^p is the learned attention coefficient for attention head p , encoding the importance of $z_{f,j}$ in the 1-hop neighborhood $N(i)$ for the update of $z_{f,i}$, and Θ^p is a learned linear transformation. α_{ii}^p represents the self-attention coefficient. Consequently, the transformation of representation $z_{f,i}$ does not solely rely on the representation itself but also receives weighted contributions from all neighboring representations $z_{f,j} \in N(i)$. This process has the potential to stabilize the predictions for patients whose initial representations may be uncharacteristic of their corresponding class but are localized within the correct data cluster in the graph. At the same time, the attention mechanism allows for less influence from patients who are wrongly clustered in the neighborhood. Moreover, this can enhance interpretability by showing which patients were considered the most in the feature refinement as highlighted in Figure 6.5. In the final step, the refined multimodal representation $z_{f,i}$ is passed to a linear layer to predict the outcome Y .

6.4 Experiments

6.4.1 Multimodal COVID-19 Datasets

KRI Dataset

The KRI dataset, an in-house dataset, comprises 132 patients, building upon the previous dataset of 65 patients reported by Burian et al. [29]. The data was collected retrospectively, adhering to the local institutional review board's guidelines and approval (ethics approval 111/20 S-KH). The patient cohort included 88 males and 44 females, ranging from 24 to 99 years of age, with an average age of 63. All patients were hospitalized at our institution between April 3rd and September 5th, 2020, and had a confirmed diagnosis of COVID-19 based on polymerase chain reaction (PCR) testing. Among the 132 patients, 53 required admission to ICU for further management. Of these ICU patients, 38 necessitated machine-assisted ventilation, and tragically, 19 patients succumbed to the disease. ICU admission criteria included the presence of at least one of the following symptoms: a respiratory rate exceeding 30 breaths per minute, peripheral oxygen saturation below 93%, an invasively

Tab. 6.1. An overview of feature extraction backbones and classifiers used for evaluation with the according modalities used as patient features and similarity metric. *Images* refer to the image features extracted with an image encoder. *Radiomics* are the radiomics calculated on the predicted segmentation. *Clinical* data encompasses vital signs, blood test results, and demographics. U-GAT is compared to a set of end-to-end methods using only clinical data (MLP-Clinical), images (ResNet18), and a U-GAT variant without auxiliary segmentation but a simple ResNet18 instead (ResNet18-GAT). Additionally, we perform experiments on image embeddings extracted from a frozen U-Net, pretrained on the same segmentation task, denoted as U-Net*. Multitasking refers to the joint optimization of classification and segmentation. [121]

Architecture	Multitasking	Multimodal	Patient modalities			Patient similarity	
			Images	Radiomics	Clinical	Radiomics	Clinical
MLP-Clinical	-	-	-	-	✓	-	-
RF-Clinical	-	-	-	-	✓	-	-
ResNet18	-	-	✓	-	-	-	-
ResNet18-GAT	-	✓	✓	-	✓	-	✓
U-Net*+RF	-	✓	-	✓	✓	-	-
U-Net*+KNN	-	✓	-	✓	✓	✓	✓
U-Net*+MLP	-	✓	✓	✓	✓	-	-
U-Net*+GraphSAGE	-	✓	✓	✓	✓	✓	✓
U-GAT*	-	✓	✓	✓	✓	✓	✓
U-GAT	✓	✓	✓	✓	✓	✓	✓

measured PaO₂/FiO₂ ratio less than 300 mmHg (1 mmHg = 0.133 kPa), respiratory failure requiring mechanical ventilation support, cardiovascular shock, or failure of various other organ systems. Upon admission to the hospital, the most common presenting symptoms were fever (66%), coughing (45%), dyspnea (33%), and gastrointestinal manifestations (15%). The mean percutaneous oxygen saturation level was $93.4 \pm 7.1\%$, and the average body temperature was $37.7 \pm 1.0^\circ C$. Notably, oxygen saturation levels differed significantly between patients admitted to the ICU and those who were not ($90.7 \pm 10.2\%$ vs. $95.0 \pm 3.5\%$), as well as between patients who required mechanical ventilation and those who did not ($89.4 \pm 10.5\%$ vs. $94.8 \pm 4.9\%$). Blood tests were performed at the time of admission, and the results, along with statistical analyses and t-test comparisons, are presented in Tables 6.2, 6.3, and 6.4.

Non-contrast, low-dose chest CT scans were performed using a 256-row MDCT scanner (iCT, Philips Healthcare, Best, The Netherlands) upon patient admission. The scans were acquired with the patient in full inspiration and arms raised. Various parameters were collected to evaluate patient outcomes, including ICU admission, need for mechanical ventilation, and survival. No distinction was made regarding whether these outcomes occurred immediately upon hospital admission or later during the patient’s stay. Expert radiologists with 4-8 years of experience manually segmented the total lung volume, healthy lung tissue, ground-glass opacities (GGO), consolidations, and pleural effusions on each CT scan. Differentiating between pleural effusion and consolidation on a voxel-wise basis was considered challenging even for senior reviewers, as both have similar Hounsfield unit ranges [128]. Moreover, pleural effusion was present in only 38 out of 132 patients and, when present, accounted for an average of 4.1% of the lung volume (1.2% for all patients). Consequently, we combined the pleural effusion and consolidation classes into a single category termed *other pathologies*. The

Tab. 6.2. KRI dataset - Blood test results upon hospital admission for the 53 ICU patients and the 79 non-ICU patients. The total number of patients (n) is less than the overall study population ($n = 132$) due to missing data for certain individuals. Statistical significance at the 5% level is indicated by *. [121]

Blood value	ICU (n=53)			No ICU (n=79)			p
	Average	Std. Dev.	n	Average	Std. Dev.	n	
Leukocytes (G/L)	8.4	4.9	53	6.7	4.1	79	0.03*
Lymphocytes (G/L)	19.3	46.9	48	22.6	35.3	75	0.65
Thrombocytes (G/L)	226.6	100.1	53	228.5	116.8	79	0.92
C-reactive protein (CRP. mg/dL)	12.19	9.30	53	6.10	6.26	78	<0.01*
Creatinine (mg/dL)	1.56	1.67	53	4.17	26.50	78	0.48
D-Dimer (μ g/mL)	5467	12801	41	1952	5570	67	0.05
Lactate dehydrogenase (LDH. U/L)	468.6	329.5	48	358.4	368.4	75	0.09
Creatinine kinase (U/L)	427.3	1167.2	48	225.3	777.2	74	0.25
Troponine-T (ng/mL)	0.071	0.161	25	0.097	0.323	34	0.71
Interleukin 6 (IL-6. pg/mL)	120.5	117.5	35	104.1	388.7	60	0.81

Tab. 6.3. KRI dataset - Blood test results at hospital admission for the 38 patients requiring ventilation and the 94 patients who did not. The total n is less than the overall study population ($n = 132$) owing to missing data for some patients. Significant differences at the 5% level are denoted by *. [121]

Blood value	Ventilation (n=38)			No Ventilation (n=94)			p
	Average	Std. Dev.	n	Average	Std. Dev.	n	
Leukocytes (G/L)	7.9	3.7	38	7.2	4.8	94	0.44
Lymphocytes (G/L)	13.1	7.9	35	24.6	46.8	88	0.15
Thrombocytes (G/L)	209.8	102.3	38	235.0	112.7	94	0.23
C-reactive protein (CRP. mg/dL)	13.54	9.80	38	6.53	6.43	93	<0.01*
Creatinine (mg/dL)	1.42	0.66	38	3.81	24.28	93	0.55
D-Dimer (μ g/mL)	5622	14484	29	2429	6019	79	0.11
Lactate dehydrogenase (LDH. U/L)	454.1	241.9	36	379.6	393.5	87	0.29
Creatinine kinase (U/L)	533.7	1354.8	35	212.7	718.0	87	0.09
Troponine-T (ng/mL)	0.049	0.089	19	0.104	0.316	40	0.47
Interleukin 6 (IL-6. pg/mL)	138.2	126.4	24	100.7	358.5	71	0.62

Tab. 6.4. KRI dataset - Blood test results upon hospital admission for the 113 surviving patients and the 19 deceased patients. The sum of n is lower than the total study population ($n = 132$) due to incomplete data for certain individuals. Statistically significant differences at the 5% level are indicated by *. [121]

Blood value	Passed (n=19)			Survived (n=113)			p
	Average	Std. Dev.	n	Average	Std. Dev.	n	
Leukocytes (G/L)	9.8	6.7	19	7.0	3.9	113	0.01*
Lymphocytes (G/L)	11.3	8.3	17	22.9	42.9	106	0.27
Thrombocytes (G/L)	201.4	99.2	19	232.2	111.6	113	0.26
C-reactive protein (CRP. mg/dL)	11.06	8.98	19	8.14	7.99	112	0.15
Creatinine (mg/dL)	1.57	0.79	19	3.37	22.13	112	0.72
D-Dimer ($\mu\text{g}/\text{mL}$)	6388	12076	13	2862	8644	95	0.19
Lactate dehydrogenase (LDH. U/L)	607.3	500.1	17	368.4	318.7	106	0.01*
Creatinine kinase (U/L)	843.2	1878.8	17	217.6	673.0	105	0.01*
Troponine-T (ng/mL)	0.120	0.216	13	0.077	0.279	46	0.61
Interleukin 6 (IL-6. pg/mL)	143.7	181.9	11	105.8	330.1	84	0.71

Tab. 6.5. KRI dataset -Radiomic features extracted from manually segmented admission CT scans. Statistical significance at the 5% level is denoted by *. [121]

Radiomic	Average	Std. Dev.	Average	Std. Dev.	p
	ICU (n=53)		No ICU (n=79)		
Healthy lung	65.2%	25.9%	92.1%	9.2%	<0.01*
GGO	22.7%	16.4%	6.2%	7.1%	<0.01*
Other pathologies	12.1%	14.0%	1.9%	4.2%	<0.01*
	Ventilation (n=38)		No Ventilation (n=94)		
Healthy lung	61.2%	22.5%	89.4%	16.1%	<0.01*
GGO	25.7%	14.9%	7.7%	10.1%	<0.01*
Other pathologies	13.1%	13.1%	3.2%	7.9%	<0.01*
	Passed (n=19)		Survived (n=113)		
Healthy lung	70.0%	20.5%	83.2%	22.0%	0.02*
GGO	22.0%	17.2%	11.1%	13.2%	<0.01*
Other pathologies	8.0%	9.7%	5.7%	10.8%	0.39

complete dataset is available upon request for research purposes within the scope of the BFS project AZ-1429-20C.

iCTCF Dataset

To demonstrate the robustness and generalizability of our approach, we have extended our evaluation to a larger, publicly available dataset: the iCTCF dataset [170] (referred to as the "external" dataset). This dataset comprises 1,521 patients and includes high-resolution CT images, clinical data, and patient outcomes. However, unlike the KRI annotations, the iCTCF dataset does not provide image annotations for different lung pathologies. As our work aims to triage COVID-19 patients, we focused on predicting the severity of outcomes in PCR-positive COVID-19 patients and excluded the control group. This resulted in a total of 894 patients, with 620 patients experiencing mild (Type I) outcomes and 274 patients experiencing severe (Type II) outcomes [170]. Due to the absence of CT image annotations in the iCTCF dataset, we employed a U-Net model, pretrained on a diverse dataset of lung CT slices [80, 202, 259] by Hofmanninger [98], to generate lung masks. Additionally, we used a nnU-Net model developed by Isensee et al. [109], which was pretrained on the COVID-19 Lung CT Lesion Segmentation Challenge dataset [201], to infer pathology annotations. Using these annotations, we extracted the radiomic feature *COVID-19 burden*, representing the percentage of the lungs affected by COVID-19-related pathologies.

6.4.2 Implementation Details

We assessed the performance of our method on the KRI dataset using a nested 5-fold cross-validation approach [3], with stratification based on ICU labels. Nested cross-validation involves two evaluation loops: an outer loop for testing and an inner loop for validation. One fold was designated as the test set in each of the five outer loops, while the remaining four folds were utilized for training and validation. Within the four inner loops, three folds were allocated for training, and one was used for validation. This process was iterated until all possible combinations were employed for testing and validation, resulting in 20 repetitions.

In the experiments presented here, following the approach of Burian et al. [29], we utilized static lung CT images acquired at the time of admission in combination with the following clinical features and blood test results: age, sex, body temperature, percutaneous oxygen saturation, leukocytes, lymphocytes, C-reactive protein (CRP), creatine, D-Dimer, lactate dehydrogenase (LDH), creatine kinase, troponin T, interleukin 6 (IL-6), and thrombocytes. The outcomes of interest were the need for mechanical ventilation, ICU admission, and patient survival (mortality), all of which were modeled as binary classification tasks. Our primary focus was on evaluating the performance of the ICU prediction task, with additional experiments conducted on the ventilation and mortality outcome tasks.

The chest CT images were sampled during the experiments using ten equidistant slices ($Z = 10$), resulting in nine subvolumes per patient. For each patient, a random subvolume was selected during the training phase. Due to the presence of only one test patient per batch, the pre-computed image features and radiomics of the other patients could be utilized. In the testing phase, a batch graph was constructed using one test node and 18 neighboring nodes

from the training set, serving as a context for the new patient. We chose concatenation as the modality aggregation function ψ for all experiments.

For the iCTCF dataset, we adopted the evaluation approach by Ning et al. [170] and split the data using a 10-fold cross-validation scheme. In each run, eight folds were allocated for training, while the remaining two were used for validation and testing. As only a single radiomic feature, the COVID-19 burden of the lung was available. We concatenated this extracted radiomic with the clinical data and did not process it separately. We encoded the resulting tabular data into a joint embedding vector of size 64 for each patient. Due to the presence of numerous features in the dataset, many of which exhibited low mutual information with the target outcome, we selected only those features with estimated mutual information higher than 0.05 for graph construction. All available clinical features were utilized as patient node features. We ended the training with early stopping after 5 epochs with no loss improvement.

U-GAT network Architecture

To accommodate the small input image size of 96×96 pixels, the initial filter size of the convolutions was reduced to 32, as opposed to the original size of 64. The model was trained using a batch size of 18 patients, with each patient represented by ten equidistant slices randomly sampled along with the corresponding clinical data. Before being passed to the classification head, the concatenated feature vector Z , comprising image, radiomics, and clinical features, underwent batch normalization and a 10% dropout. We chose not to backpropagate the classification loss through the extracted radiomics R over the U-Net output, as it significantly deteriorated the image segmentation performance. We evaluated three approaches for feature fusion Ψ : concatenation, averaging, and max pooling. Based on the performance evaluation conducted on the validation set, concatenation yielded marginally better results than the other two approaches, albeit without a statistically significant difference. Consequently, we adopted the concatenation approach for the experiments performed in this study. See Section 3.3.1 for more details on pooling operations.

The graph-based classification head consists of a Graph Attention Network (GAT) with two layers, five attention heads, and a dropout rate of 10%. In the first layer, each node feature vector, which has an input size of 96, is refined to a feature size of 64. The final node classification layer further reduces the feature size to match the number of classification labels. For all binary classification outputs, a sigmoid activation function is applied.

For the segmentation and image feature extraction backbone, we opted for the classical 2D U-Net architecture proposed by Ronneberger et al.[199], with several modifications to the double convolution blocks. We added a batch normalization layer after each activation to facilitate faster convergence. Additionally, we applied one-pixel padding in each convolution layer to ensure alignment between the network's input and output image sizes. The final layer of the U-Net consisted of a one-dimensional convolution, which reduced the feature maps to the number of output classes, followed by a softmax layer. To train the segmentation network, we employed the Dice loss introduced by Milletari[163], while a binary cross-entropy (BCE) loss was used for the classification task.

Graph Construction

Drawing inspiration from the work of Parisot et al. [179], we define the similarity $\text{Sim}(u, v)$ between two nodes u and v in the graph by applying a radial basis function (RBF) kernel to their distance. The RBF kernel is parameterized by the mean distance μ , which is calculated over the training set to ensure a data-driven approach to capturing the underlying structure of the graph

$$\text{Sim}(u, v) = \exp\left(-\frac{\omega(u, v)}{2\mu^2}\right). \quad (6.3)$$

We constructed the KNN graph using the mutual information-weighted distance metric introduced in Section 6.3.1 for the following experiments. This method was selected after comparing its performance to other graph construction techniques on the validation set. The weighted Euclidean distance (Minkowski distance of order $p = 2$) was chosen as the distance metric ω , with each feature dimension weighted by its estimated mutual information with the corresponding outcome label. The mutual information was estimated using the approach proposed by Ross et al. [200], which involves averaging the results of 30 repetitions using three nearest neighbors. We compared the mutual information-weighted KNN graph against a Pearson correlation-weighted KNN and an unweighted KNN to assess the impact of feature weighting. Different manually selected feature subsets were evaluated for the unweighted setup, as shown in Table 6.8. A hyperparameter search on the validation set determined the optimal number of neighbors k used in the graph construction process.

Training Details

All experiments were performed using PyTorch 1.7.0 [180] and PyTorch Geometric 1.7.0 [74], with the Adam optimizer configured with a base learning rate of 5×10^{-4} and a weight decay of 3×10^{-5} . The models were trained on an NVIDIA Titan V 12GB GPU, utilizing Polyaxon for the KRI dataset. An epoch was defined as 80 patients, and the training duration was set to a minimum of 25 epochs for end-to-end cases and 5 epochs when using a pretrained U-Net, denoted as U-Net, in the experiments. The training was terminated if the validation loss did not improve for five consecutive epochs after reaching the minimum number of epochs. In the end-to-end experiments involving joint segmentation and classification, a pretraining schedule was employed based on its observed benefits in preliminary experiments. The classification loss was set to zero for the initial 20 epochs, while the segmentation loss was trained solely on the lung masks for the first 10 epochs and then on all segmentation labels for an additional 10 epochs. For estimating mutual information and constructing the KNN graph, the scikit-learn library 0.24.1 [183] was utilized. Correlation calculations were performed using SciPy 1.6.2 [238], and NumPy 1.18.2 [91] was employed for all distance calculations. Using default parameters, the Random Forest implementation was based on scikit-learn 0.24.1 [183].

6.4.3 Ablative Testing and Baselines

To assess the impact of the various components in our method, we present ablative results on the test set, focusing on two main aspects: the image and radiomics feature extraction performed by the U-Net and the GAT classification. We compare the end-to-end U-GAT feature extraction with features obtained from a simple frozen U-Net trained on the same annotations without multi-tasking and the end-to-end image features from a ResNet18 architecture, as

introduced by He et al. [92]. It is crucial to note that radiomics were not incorporated into the ResNet18-GAT architecture since ResNet18 does not generate segmentations. To evaluate the effectiveness of GAT, we compare it with the following alternative classification methods: Here is the rewritten itemized list:

- **Weighted K-nearest neighbors (KNN):** A standard weighted k-nearest neighbor classifier from the scikit-learn library, which employs the inverse Euclidean distance of all features as the similarity metric for selecting neighbors and weighting their labels [183].
- **Multilayer Perceptron (MLP):** A simple neural network classifier consisting of a hidden layer with 64 units, followed by a leaky ReLU activation function and a 10% dropout rate.
- **GraphSAGE:** A variant of our method that replaces the GAT operator with GraphSAGE, a similar Graph Convolutional Network (GCN) that does not incorporate an attention mechanism, as proposed by Hamilton et al. [90].

To investigate the benefits of multimodal learning further, we compare the performance of unimodal and multimodal approaches by evaluating an MLP classifier using either clinical data or image features extracted by a ResNet18 only. Table 6.1 summarizes the data types utilized in each method, facilitating a comprehensive understanding of the input modalities and their impact on the classification results.

U-GAT Ensemble and Random Forest

Random Forest, an ensemble method, is an effective classifier for small datasets due to its robustness against overfitting, as discussed in Section 3.2.4. Additionally, Random Forests offer the benefit of interpretability, making them an attractive choice for many applications. Previous studies by Burian et al. [29] and Chao et al. [40] have successfully employed Random Forests to utilize tabular radiomics and clinical data for predicting ICU admission, as discussed in Section 6.2.1. In this experiment, we focus on the task of ICU prediction and investigate whether an ensemble of our proposed model can enhance its performance by increasing its resilience to overfitting. We also compare the performance of our ensemble approach to the well-established Random Forest classifier. To create an ensemble, we average the predicted probabilities of the four models trained on the inner loops of the nested cross-validation and evaluate their performance on the five test sets of the outer loop of the nested cross-validation. This approach allows us to assess the generalization capabilities of the ensemble and compare it to the individual models and the Random Forest classifier.

6.4.4 Metrics

The primary metrics for evaluating the binary classification of outcome predictions are average precision (AP) and the area under the receiver operating characteristic curve (AUC). These metrics are chosen due to their independence from specific classification thresholds, providing a comprehensive assessment of the model's discriminative power. Considering the severe class imbalance present in all tasks, the F1 score (F1) is selected as the main threshold-dependent metric. Additionally, the balanced accuracy score (bACC), sensitivity,

and specificity are reported in the ensemble experiments to provide a more comprehensive view of the model's performance. For all threshold-dependent metrics, the optimal threshold is determined using the validation results and maximizing Youden's J statistic [263], calculated as $J = \text{sensitivity} + \text{specificity} - 1$. All binary classification metrics are computed using scikit-learn 0.24.1 [183].

To evaluate the segmentation performance of our multitasking approach, we quantify the overlap between segmented regions and ground truth using the Dice score (DS), which assesses the spatial agreement between the predicted and actual segmentations.

6.5 Results and Discussion

6.5.1 Population Graph Construction

In the first set of experiments on our KRI dataset, we optimized the population graph construction method. This involved evaluating various feature selections and distance weights to improve the KNN-based graph construction. We found that connecting each node with its seven nearest neighbors provided optimal results based on a hyperparameter search using a simple, unweighted KNN classifier. To weight features in the distance calculation of the similarity metric used for KNN neighbor selection, we employed two measures: mutual information and Pearson correlation. Table 6.6 presents the top 10 features based on the average of both measures for the ICU task. While some features exhibited a Pearson correlation > 0.3 and mutual information > 0.1 in the ICU and ventilation tasks, the mortality task demonstrated significantly lower values, highlighting the inherent difficulty of the task. Across all tasks, the percentage of healthy lung tissue displayed the highest mutual information. The results in Table 6.8 confirmed that our proposed mutual information-based weighting method achieved the best performance, particularly for the ICU task, with an AP of 0.722 ± 0.096 and an AUC of 0.757 ± 0.142 . Comparing our approach with manual feature selection, such as using only clinical data, revealed that incorporating all available features is most effective. However, estimating mutual information can further assist in identifying the most relevant features and assigning them higher weights in the similarity metric. The external dataset reinforced the significance of radiomic data, with the COVID-19 burden exhibiting the highest mutual information with the severity labels (see Table 6.7). One of the key advantages of employing a weighted distance for KNN graph construction is its adaptability to each task without requiring prior knowledge. Figure 6.2 illustrates the graph for each task on the KRI dataset, both with and without weighting the distance measure using mutual information. In addition to improving classification performance, an effective similarity measure can be utilized to identify relevant patients who have been treated in the past, supporting physicians' decision-making process by enabling them to analyze the disease progression in them.

6.5.2 U-GAT Evaluation

Table 6.9 presents the results of our next set of experiments, which evaluate the different components of the proposed method and compare them to baseline approaches. Our multi-

Tab. 6.6. KRI dataset - Top 10 features ranked by their mutual information and Pearson correlation for each task, calculated as the average across the training sets of all repetitions. For the multilabel setup, the mutual information between each feature and the ordinal regression of outcome severity is estimated. [121]

Task	Feature	Category	Mutual information	Pearson correlation
ICU	Healthy lung (%)	Radiomics	0.244 ± 0.052	-0.596 ± 0.033
ICU	Ground-glass opacity (%)	Radiomics	0.184 ± 0.043	$+0.577 \pm 0.026$
ICU	Other pathologies (%)	Radiomics	0.144 ± 0.055	$+0.471 \pm 0.048$
ICU	C-reactive protein	Clinical	0.104 ± 0.038	$+0.372 \pm 0.071$
ICU	Interleukin 6	Clinical	0.091 ± 0.023	$+0.091 \pm 0.137$
ICU	Age	Clinical	0.087 ± 0.031	$+0.018 \pm 0.062$
ICU	Lymphocytes	Clinical	0.047 ± 0.027	-0.062 ± 0.112
ICU	Temperature	Clinical	0.043 ± 0.040	-0.016 ± 0.116
ICU	Serum creatinine	Clinical	0.041 ± 0.045	$+0.009 \pm 0.125$
ICU	Thrombocytes	Clinical	0.039 ± 0.037	-0.007 ± 0.060
ICU	Creatine kinase (total)	Clinical	0.037 ± 0.040	$+0.113 \pm 0.110$
Ventilation	Healthy lung (%)	Radiomics	0.212 ± 0.033	-0.581 ± 0.030
Ventilation	Ground-glass opacity (%)	Radiomics	0.170 ± 0.022	$+0.585 \pm 0.026$
Ventilation	Other pathologies (%)	Radiomics	0.159 ± 0.055	$+0.428 \pm 0.051$
Ventilation	Interleukin 6	Clinical	0.114 ± 0.048	$+0.109 \pm 0.130$
Ventilation	C-reactive protein	Clinical	0.082 ± 0.047	$+0.395 \pm 0.070$
Ventilation	Temperature	Clinical	0.082 ± 0.044	$+0.031 \pm 0.118$
Ventilation	Age	Clinical	0.059 ± 0.037	$+0.056 \pm 0.053$
Ventilation	Serum creatinine	Clinical	0.055 ± 0.034	-0.020 ± 0.063
Ventilation	Lactate dehydrogenase	Clinical	0.053 ± 0.028	$+0.104 \pm 0.060$
Ventilation	Percutaneous oxygen saturation	Clinical	0.052 ± 0.017	-0.285 ± 0.074
Ventilation	Creatine kinase (total)	Clinical	0.045 ± 0.046	$+0.160 \pm 0.106$
Mortality	Healthy lung (%)	Radiomics	0.061 ± 0.040	-0.210 ± 0.093
Mortality	C-reactive protein	Clinical	0.048 ± 0.034	$+0.126 \pm 0.072$
Mortality	Lymphocytes	Clinical	0.034 ± 0.040	-0.095 ± 0.030
Mortality	Percutaneous oxygen saturation	Clinical	0.033 ± 0.038	-0.023 ± 0.068
Mortality	Interleukin 6	Clinical	0.031 ± 0.013	$+0.068 \pm 0.096$
Mortality	D-dimer	Clinical	0.030 ± 0.023	$+0.122 \pm 0.117$
Mortality	Temperature	Clinical	0.022 ± 0.026	-0.014 ± 0.068
Mortality	Lactate dehydrogenase	Clinical	0.019 ± 0.024	$+0.246 \pm 0.070$
Mortality	Sex	Clinical	0.019 ± 0.010	-0.150 ± 0.041
Mortality	Ground-glass opacity (%)	Radiomics	0.018 ± 0.023	$+0.265 \pm 0.079$
Mortality	Other pathologies (%)	Radiomics	0.016 ± 0.024	$+0.083 \pm 0.100$
Multilabel	Healthy lung (%)	Radiomics	0.274 ± 0.063	-0.548 ± 0.051
Multilabel	Ground-glass opacity (%)	Radiomics	0.190 ± 0.052	$+0.550 \pm 0.042$
Multilabel	Other pathologies (%)	Radiomics	0.173 ± 0.057	$+0.407 \pm 0.066$
Multilabel	Interleukin 6	Clinical	0.105 ± 0.040	$+0.098 \pm 0.133$
Multilabel	Sex	Clinical	0.104 ± 0.110	-0.167 ± 0.052
Multilabel	C-reactive protein	Clinical	0.100 ± 0.043	$+0.352 \pm 0.068$
Multilabel	Lymphocytes	Clinical	0.062 ± 0.044	-0.116 ± 0.046
Multilabel	Age	Clinical	0.057 ± 0.028	$+0.057 \pm 0.062$
Multilabel	Percutaneous oxygen saturation	Clinical	0.047 ± 0.055	-0.233 ± 0.069
Multilabel	Troponin T	Clinical	0.040 ± 0.021	$+0.068 \pm 0.103$
Multilabel	Temperature	Clinical	0.036 ± 0.038	-0.015 ± 0.108

Tab. 6.7. iCTCF dataset - Top 10 features ranked by their mutual information with the outcome severity (Type I vs. Type II) and Pearson correlation, averaged across the training sets of all ten folds. The overall mutual information and Pearson correlation values are lower than the tasks in our in-house dataset. The COVID-19 burden radiomic feature, extracted from the U-Net and equivalent to one minus the healthy lung percentage, consistently ranks among the most important features. [121]

Task	Feature	Category	Mutual information	Pearson correlation
iCTCF Severity	Neutrophil percentage (NEP)	Clinical	0.074 ± 0.011	0.315 ± 0.016
iCTCF Severity	COVID-19 burden	Radiomic	0.067 ± 0.011	0.376 ± 0.021
iCTCF Severity	Lymphocyte percentage (LYP)	Clinical	0.066 ± 0.014	-0.293 ± 0.017
iCTCF Severity	Lymphocyte count (LY)	Clinical	0.066 ± 0.018	-0.229 ± 0.029
iCTCF Severity	Prothrombin time (PT)	Clinical	0.045 ± 0.013	0.063 ± 0.011
iCTCF Severity	Calcium (CA)	Clinical	0.043 ± 0.008	-0.242 ± 0.015
iCTCF Severity	D-dimer (DD)	Clinical	0.032 ± 0.008	0.225 ± 0.015
iCTCF Severity	Albumin (ALB)	Clinical	0.032 ± 0.006	-0.269 ± 0.021
iCTCF Severity	Basophil percent (BAP)	Clinical	0.030 ± 0.007	-0.062 ± 0.011
iCTCF Severity	Neutrophil count (NE)	Clinical	0.030 ± 0.016	0.220 ± 0.018

Tab. 6.8. Comparison of edge features and their weighting schemes for distance calculation, evaluated on the validation set of the KRI dataset. [121]

Task	Architecture	Distance features	Distance feature weights	AP	AUC
ICU	U-GAT*	age, sex	-	0.512 ± 0.109	0.573 ± 0.109
ICU	U-GAT*	clinical	-	0.671 ± 0.152	0.720 ± 0.135
ICU	U-GAT*	radiomics	-	0.670 ± 0.145	0.720 ± 0.116
ICU	U-GAT*	all	-	0.704 ± 0.080	0.733 ± 0.073
ICU	U-GAT*	all	Pearson correlation	0.697 ± 0.122	0.751 ± 0.088
ICU	U-GAT*	all	Mutual information	0.722 ± 0.096	0.757 ± 0.142

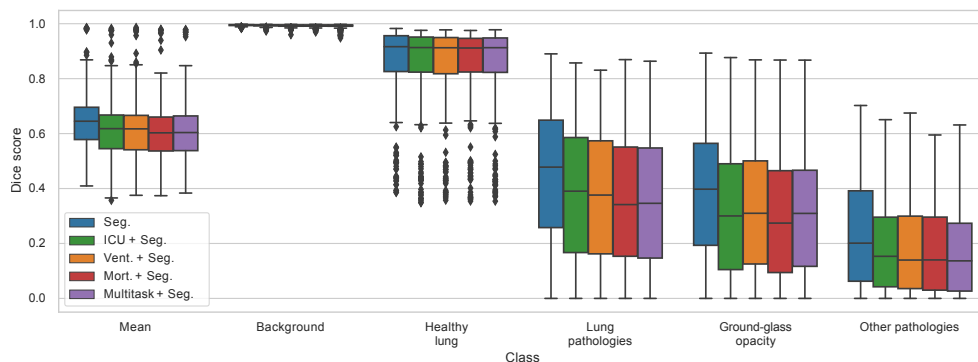


Fig. 6.3. Boxplot visualization of the Dice scores for the segmentation results of different methods on the KRI dataset. The multitasking approaches are compared with single-task segmentation. Although the auxiliary segmentation task improves classification results, the segmentation performance is lower in all multitask settings compared to the U-Net optimized solely for segmentation. [121]

Tab. 6.9. Ablative testing and comparison of the proposed method with an MLP using only clinical data and a ResNet18 using only image data as input for all tasks. U-GAT* denotes the proposed method using image and radiomic features extracted from a frozen U-Net trained on the same annotations as the end-to-end U-GAT. Values marked with † indicate statistical significance ($p < 0.05$) based on Wilcoxon’s rank test comparing the proposed method with each baseline. [121]

Dataset	Task	Architecture	AP	AUC	F1
KRI	ICU	MLP-Clinical	$0.577 \pm 0.109^\dagger$	$0.654 \pm 0.104^\dagger$	$0.560 \pm 0.107^\dagger$
KRI	ICU	ResNet18	0.670 ± 0.097	0.716 ± 0.077	$0.560 \pm 0.084^\dagger$
KRI	ICU	U-Net*+KNN	0.632 ± 0.113	0.677 ± 0.112	$0.519 \pm 0.131^\dagger$
KRI	ICU	U-Net*+MLP	$0.615 \pm 0.127^\dagger$	0.687 ± 0.128	0.612 ± 0.085
KRI	ICU	U-Net*+GraphSAGE	$0.628 \pm 0.114^\dagger$	$0.690 \pm 0.107^\dagger$	$0.574 \pm 0.085^\dagger$
KRI	ICU	ResNet18-GAT	0.637 ± 0.165	0.678 ± 0.160	$0.595 \pm 0.084^\dagger$
KRI	ICU	U-GAT*	0.672 ± 0.129	0.725 ± 0.107	0.651 ± 0.104
KRI	ICU + Seg.	U-GAT	0.699 ± 0.149	0.743 ± 0.103	0.661 ± 0.084
KRI	Ventilation	MLP-Clinical	0.527 ± 0.167	$0.692 \pm 0.109^\dagger$	0.475 ± 0.188
KRI	Ventilation	ResNet18	0.573 ± 0.127	$0.715 \pm 0.086^\dagger$	$0.390 \pm 0.160^\dagger$
KRI	Ventilation	U-Net*+KNN	$0.527 \pm 0.180^\dagger$	$0.674 \pm 0.112^\dagger$	$0.368 \pm 0.192^\dagger$
KRI	Ventilation	U-Net*+MLP	0.587 ± 0.183	0.741 ± 0.119	0.488 ± 0.134
KRI	Ventilation	U-Net*+GraphSAGE	0.603 ± 0.151	0.758 ± 0.109	0.481 ± 0.205
KRI	Ventilation	ResNet18-GAT	0.570 ± 0.152	$0.689 \pm 0.152^\dagger$	$0.423 \pm 0.178^\dagger$
KRI	Ventilation	U-GAT*	0.618 ± 0.137	0.788 ± 0.106	0.592 ± 0.130
KRI	Vent. + Seg.	U-GAT	0.644 ± 0.142	0.788 ± 0.112	0.539 ± 0.179
KRI	Mortality	MLP-Clinical	0.261 ± 0.135	0.544 ± 0.134	0.224 ± 0.152
KRI	Mortality	ResNet18	$0.210 \pm 0.116^\dagger$	$0.461 \pm 0.155^\dagger$	0.155 ± 0.138
KRI	Mortality	U-Net*+KNN	0.257 ± 0.137	0.512 ± 0.166	0.184 ± 0.147
KRI	Mortality	U-Net*+MLP	0.252 ± 0.157	0.502 ± 0.191	0.190 ± 0.157
KRI	Mortality	U-Net*+GraphSAGE	0.270 ± 0.143	0.568 ± 0.180	0.236 ± 0.163
KRI	Mortality	ResNet18-GAT	0.247 ± 0.151	0.520 ± 0.156	0.184 ± 0.157
KRI	Mortality	U-GAT*	0.271 ± 0.137	0.549 ± 0.188	0.230 ± 0.172
KRI	Mort. + Seg.	U-GAT	0.287 ± 0.186	0.586 ± 0.187	0.199 ± 0.173
iCTCF	Severity	MLP-Clinical	0.556 ± 0.099	0.735 ± 0.068	0.539 ± 0.064
iCTCF	Severity	ResNet18	0.525 ± 0.140	0.739 ± 0.083	0.513 ± 0.102
iCTCF	Severity	U-Net*+KNN	$0.456 \pm 0.070^\dagger$	0.705 ± 0.060	$0.318 \pm 0.129^\dagger$
iCTCF	Severity	U-GAT*	0.558 ± 0.102	0.740 ± 0.096	0.505 ± 0.114
iCTCF	Severity	U-GAT	0.593 ± 0.106	0.763 ± 0.085	0.521 ± 0.109

Tab. 6.10. iCTCF Dataset - Comparison of test set DICE scores between U-Net and U-GAT on the iCTCF dataset. The joint optimization of segmentation and classification leads to a minor decrease in segmentation metrics. [121]

Architecture	Segmentation	Classification	Dice lung	Dice COVID-19
U-Net	✓	-	0.984 ± 0.002	0.738 ± 0.019
U-GAT	✓	✓	0.970 ± 0.038	0.718 ± 0.027

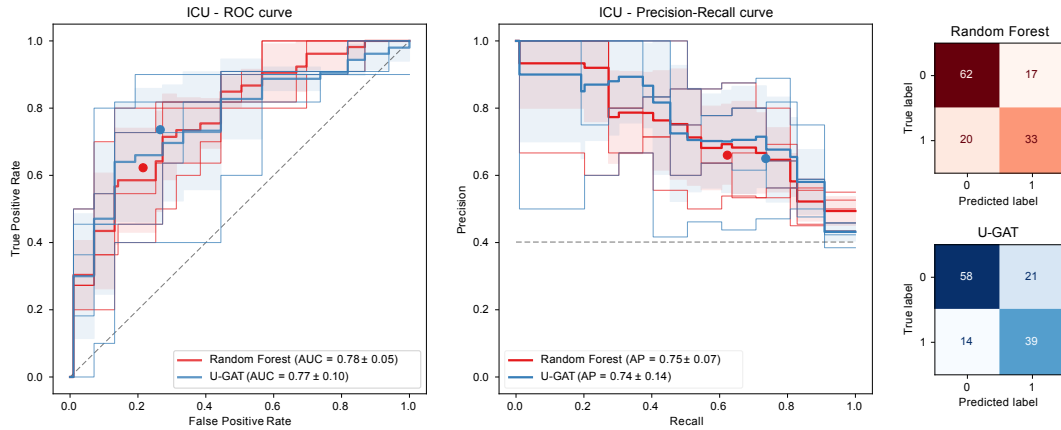


Fig. 6.4. KRI dataset - Ensemble results for the ICU task on the KRI dataset: receiver operating characteristic (ROC) and precision-recall curves comparing inner cross-validation loop ensembles of Random Forest and the proposed U-GAT method. Each narrow curve represents the results of one of the 5 test sets. The average of these curves is estimated with the bold curve, and the shaded area depicts its standard deviation. The confusion matrices on the right include the predictions for all patients across all 5 test folds with optimized thresholds. The corresponding metrics of these confusion matrices are visualized with filled circles in the curve diagrams.

modal method outperforms the unimodal MLP, which is limited to using only clinical data as input. A similar trend is observed when restricting the model to solely utilize imaging data, as is the case for the ResNet18 method. In this scenario, our proposed methods consistently outperform ResNet18 across all tasks, highlighting the advantages of a multimodal approach. U-GAT achieves a higher AP than the other methods in all ablations, including replacing the U-Net with a ResNet18 and replacing the GAT with an MLP or a GraphSAGE. These results demonstrate the value of leveraging similar patients from the training set to refine the features of test patients. Comparable findings are observed on the external dataset, where U-GAT attains a higher AP of 0.593 ± 0.106 compared to the single modality models MLP and ResNet18, which achieve 0.556 ± 0.099 and 0.525 ± 0.140 , respectively, further emphasizing the benefits of multimodal learning.

The joint end-to-end training of the segmentation and classification tasks appears to slightly improve the AP for all tasks. However, as shown in Figure 6.3, the average Dice score is lower in all multitask setups compared to the segmentation single-task setup. In this configuration, both the ICU prediction and the ventilation prediction reached their highest AP values of 0.699 ± 0.149 and 0.644 ± 0.142 , respectively. The end-to-end multitasking of the classification of all labels and segmentation only benefited the mortality task, achieving the highest AP of 0.289 ± 0.138 and AUC of 0.620 ± 0.175 in this setup (see Table 6.11).

Tab. 6.11. KRI dataset - Results for the multitasking of pathology segmentation and the prediction of three patient outcomes (ICU admission, need for ventilation, and mortality) on the KRI dataset. The graph construction is based on the ordinal regression of outcome severity. Each outcome prediction is modeled as a non-exclusive binary classification, i.e., a multilabel problem. The mortality task is the only task benefiting from this multitask setup. [121]

Task	Architecture	AP	AUC	bACC**	F1**
ICU	U-GAT	0.649 ± 0.128	0.697 ± 0.116	0.642 ± 0.097	0.569 ± 0.163
Ventilation	U-GAT	0.622 ± 0.127	0.774 ± 0.094	0.681 ± 0.102	0.503 ± 0.188
Mortality	U-GAT	0.289 ± 0.138	0.620 ± 0.175	0.536 ± 0.133	0.216 ± 0.174

Tab. 6.12. Comparative analysis of ICU outcome prediction on the KRI dataset: U-GAT versus its cross-validation ensemble, a random forest model using only clinical data, and another random forest model incorporating all available tabular data, including radiomics extracted with a pretrained U-Net. [121]

Architecture	AP	AUC	bACC	F1	Sens.	Spec.
RF-Clinical	0.635 ± 0.098	0.707 ± 0.086	0.624 ± 0.056	0.519 ± 0.070	0.475 ± 0.131	0.773 ± 0.175
U-Net*+RF	0.729 ± 0.089	0.774 ± 0.057	0.716 ± 0.075	0.649 ± 0.011	0.651 ± 0.177	0.781 ± 0.166
U-GAT ensemble	0.745 ± 0.137	0.770 ± 0.098	0.735 ± 0.111	0.700 ± 0.114	0.736 ± 0.067	0.734 ± 0.174

The mortality task generally yields worse results, which can be primarily attributed to the severe data imbalance present for this task, with only 19 out of 132 positive samples. Furthermore, we observe low mutual information between the radiomics and clinical features and the mortality outcome (see Table 6.6), suggesting that the features might not be sufficiently predictive for this specific task. Several relevant clinical aspects closely related to multiorgan failure, such as heart, kidney, and liver parameters, were unavailable in the datasets. The evaluation on the external dataset reveals a similar trend, where joint end-to-end training of severity classification and pathology segmentation with U-GAT increases the AP from 0.558 ± 0.102 to 0.593 ± 0.106 compared to U-GAT*, which uses segmentations from a frozen U-Net trained on the same annotations.

Multitasking Evaluation

To investigate the potential synergies between segmentation and classification tasks, as well as the simultaneous prediction of different patient outcomes, we conducted additional experiments exploring the interdependence of these tasks. The results presented in Table 6.11 demonstrate that joint segmentation can benefit classification performance. However, only the mortality prediction task exhibited improvement among the patient outcomes when all outcomes were predicted concurrently. These findings suggest that while multitasking can offer advantages in certain scenarios, the specific combination of tasks and their interrelationships play a crucial role in determining the extent of the benefits observed.

U-GAT Ensemble and Random Forest

In line with the discussion in Section 6.4.3, we compare our method against Random Forests, which have been employed in previous works to perform classification by fusing tabular radiomics with clinical data. The comparison, presented in Table 6.12, highlights the improvement in U-GAT's average precision from 0.699 ± 0.149 to 0.745 ± 0.137 when an ensemble

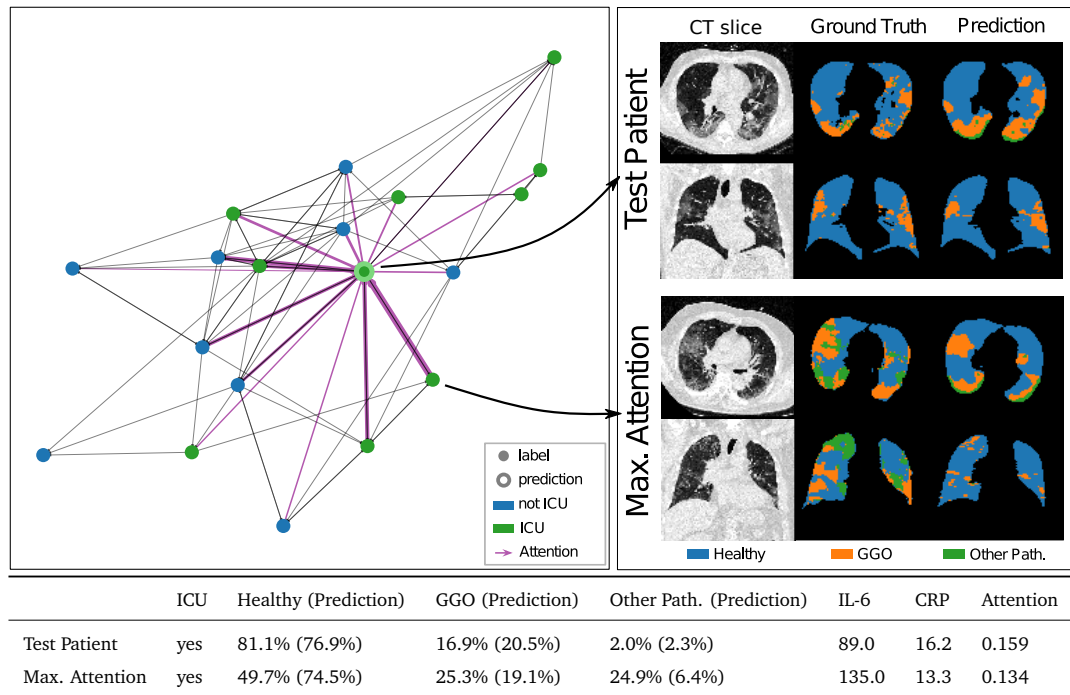


Fig. 6.5. Visualization of the neighborhood attention of GAT for a single test patient from the KRI dataset. Left: Batch graph showing the attention scores of the test patient’s neighbors after two hops. The line thickness corresponds to the attention score of each neighbor. Right: CT images, segmentation ground truth, and predicted segmentation of a single axial and coronal slice from the test patient and its neighbor with the maximum attention score. Bottom: Most important features for the test patient and the neighbor with maximum attention, with the radiomics predicted by the pretrained U-Net shown in brackets. [121]

approach is used. This enhancement allows U-GAT to slightly outperform the Random Forest, which achieves an average precision of 0.729 ± 0.089 . These results suggest that ensembling our method increases its resilience to overfitting, yielding performance comparable to that of a Random Forest. The AP and AUC metrics are higher in the respective Random Forest models. However, the U-GAT ensemble exhibits the highest balanced accuracy, F1 score, and sensitivity, while the Random Forest ensemble demonstrates the highest specificity. The standard deviations and the receiver operating characteristic (ROC) and precision-recall curves in Figure 6.4 reveal that although the averages of both ensembles are similar, the U-GAT ensemble has a wider spread with outliers in both directions. A Wilcoxon signed-rank test did not indicate a significant difference in AP or AUC between the two ensembles.

6.5.3 Interpretability and Graph Attention

In addition to its performance advantage over GraphSAGE, using GAT in our model offers another significant benefit. The attention mechanism learns to identify the most relevant neighbors in the graph for the prediction task, providing valuable insights into the model’s decision-making process. By analyzing the attention scores, we can gain a better understanding of the patients that the model considers most influential for individual outcome predictions.

The connections within the patient population graph can help uncover new information about poorly understood diseases and provide valuable insights to physicians. When combined with

the segmentation results, our attention mechanism allows clinicians to inspect the model's output and decision-making process, potentially increasing their confidence in the predictions. For each of the two GAT layers, the model assigns attention scores to the neighbors of each node in the graph. These scores determine the extent to which the node representation after the layer will be based on the representations of its different one-hop neighbors. The attention scores can be conceptualized as a weighted directed adjacency matrix $A \in [0, 1]^{N \times N}$, where N is the number of nodes in the batch and all rows in A sum to 1. By multiplying the attention matrices of both layers, we obtain a matrix that shows how the representation of a node is influenced by its two-hop neighborhood, i.e., all nodes that are at most two edges away.

The qualitative results for the test patient shown in Figure 6.5 demonstrate that the attention mechanism successfully assigns high importance to neighbors of the same class and lower importance to those of the opposite class, effectively refining the neighborhood constructed by the KNN algorithm. Moreover, we observe that the attention mechanism does not necessarily assign high attention to neighbors particularly similar in their radiomic or clinical features. Unlike a simple KNN classifier, which can only base its predictions on feature similarity, our method can identify the most relevant neighbors that go beyond simple correlations and are connected through more complex patterns. This ability to capture orthogonal information embedded in the KNN graph sets our approach apart and contributes to its improved performance.

6.5.4 Challenges and Outlook

There is room for improvement in our model's capability of segmenting infrequent lung pathologies, such as pleural effusion, and predicting imbalanced outcomes, particularly mortality. While weighting the patient distance metric with mutual information effectively improved the graph construction and the elevated features that align with established radiological findings, it is important to acknowledge the remaining challenges. The mutual information exhibits a substantial standard deviation throughout tasks and is notably lower for the mortality prediction task. This observation highlights the inherent complexity of predicting mortality and suggests a potential lack of informative features given the available features in the dataset. These challenges can be attributed to the limited size of the patient cohort and the scarcity of annotated data for certain pathologies and outcomes. Future studies should focus on expanding the patient cohort and incorporating more comprehensive clinical data to address these limitations.

6.6 Conclusion

We propose U-GAT, an end-to-end methodology that integrates CT lung scans, clinical data, and radiomics to form a multimodal patient graph for outcome prediction in COVID-19 patients. The Graph Attention Network (GAT) processes this graph, stabilizing and supporting predictions based on similar patients. Automatically extracted radiomics from the generated segmentation improve performance over baselines, and the auxiliary segmentation of COVID-19 pathologies enhances outcome prediction accuracy. Our novel feature weighting using mutual information enables task-specific patient clustering, allowing the model to learn from

similar cases. The GAT's attention mechanism provides interpretability, giving clinicians insights into the model's decision-making process.

U-GAT mirrors the clinical reasoning process, sharing similarities with the exemplar theory of clinical reasoning (Section 2.1.3). The patient representations and attention mechanisms are analogous to how clinicians retrieve and focus on relevant exemplars from memory when faced with a new case. Visualizing the considered patient neighborhood and intermediate image information like segmentation output and radiomics for both the evaluated and related patients gives insights into the model's reasoning and can provide helpful information for clinicians. In particular, a detailed view of most similar previous patients with the highest attention, as shown in Figure 6.5, could retrieve relevant information for clinicians offering a pathway towards clinical reasoning support.

Integrating imaging, clinical data, and radiomics aligns with integrated diagnostics (Section 2.1), capturing the complex relationships and patterns clinicians synthesize during diagnosis. However, developing accurate, transparent, and robust predictive models for complex clinical scenarios remains challenging, particularly with limited data. While U-GAT emulates certain aspects of clinical reasoning, further research must refine and validate this methodology in real-world settings and with data at scale. Future work should expand the patient cohort, incorporate more comprehensive clinical data, and explore techniques for handling imbalanced data and infrequent pathologies.

Part III

Cross-modal Extraction of Structured
Knowledge

Structured and Unstructured Clinical Knowledge

The previous Chapter 4 explored using experiential knowledge for clinical decision support with deep learning, focusing on how multimodal patient information can be used to model relationships in a population graph and how this graph can subsequently be processed for decision-making. This part shifts focus to the types of knowledge available about individual patients and how this knowledge can be tested against hypotheses formed using prior knowledge in clinical reasoning support, particularly radiological decision-making.

In Chapter 8, we investigate how large amounts of unstructured but paired multimodal data can be used to train a contrastive pre-training model (see Section 3.3.3), enabling the extraction of structured findings from chest X-rays with only a few training examples of filled structured reporting templates.

Building upon this work, Chapter 9 extends the method using a classification-by-description approach in a zero-shot fashion that does not require any labeled samples. Here, we test hypotheses of predefined findings in chest X-rays by estimating the probability of radiological observations associated with each finding according to prior knowledge. The approach resembles an analytical top-down method while accounting for the uncertainties of compositional image observations.

As seen in Section 3.1, various structured patient information can be used in clinical decision-making, such as demographics, vital signs, blood lab results, standardized diagnosis codes, radiological findings, and patient outcomes. The availability of well-structured information is crucial, as it forms the basis for retrospective clinical trials, generating new insights for evidence-based medicine and serving as a supervision signal for machine learning models. However, a significant challenge for data-driven systems is that a lot of data is not structured in this way, and even when structured, it may not be standardized or aggregatable due to interoperability issues [193]. A similar distinction applies to formal knowledge (Section 2.1.4), which can be stored as unstructured text or in structured formats like knowledge graphs and databases.

7.1 Standardization and Structured Reporting

Various standardization systems have been introduced in medicine at both national and international levels to address the challenges of unstructured and non-standardized data. The most common method of radiology reporting involves typing or dictating an unstructured free-text report. While this approach requires no additional tools, it can be time-consuming, and the lack of standardization hinders retrospective analysis and the use of reported findings

for machine learning. In contrast, structured reporting involves filling structured reports that are either standardized based on consensus within the medical community or can be easily matched to standardized reports. [66, 76]

The Radiological Society of North America (RSNA) has developed several standards to promote structured reporting and interoperability [130]. These include:

- **RadLex**¹, a comprehensive lexicon incorporating radiology-specific terms such as anatomy, diseases, and imaging findings;
- **RadElement**², a framework for Common Data Elements (CDEs) that provides a standardized way to define report elements, promoting consistency and facilitating research; and
- **RadReport**³, a web-based library offering best-practice report templates encoded using the MRRT profile.

MRRT, which stands for Management of Radiology Report Templates, has been developed by the Integrating the Healthcare Enterprise (IHE) committee to enable the interoperability of reporting templates across organizations and countries. These efforts have been joined by the European Society of Radiology (ESR) and the German Society of Radiologists (DRG), which has also started providing templates for various clinical applications⁴, similar to the American RadReport website [132].

7.1.1 Structured Reporting in Deep Learning

Standardized reports can be modeled in various ways for deep learning. Beyond the simple one-hot encoding of clinical findings [190, 222] that do not capture any dependencies, as discussed in Section 8.1.1, various graph-based approaches such as RadGraph [110] and ImaGenome [253] have been proposed. These approaches model the content of reports with a fixed set of nodes and relationships between them, describing the presence, location, and attributes of radiological findings found in chest X-ray reports.

In FLEXR, presented in Chapter 8, we predict triplets extracted from ImaGenome resembling granular findings that could be part of a structured reporting template. Additionally, we perform experiments using a real-life structured reporting template from RadReport on the severity of cardiomegaly. In our work Prior-RadGraphFormer [254], we propose an image-to-graph model that integrates a knowledge graph containing all possible graph combinations to generate a patient-specific radiology graph that could be used to either fill a structured reporting template or generate an unstructured free-text report.

Another approach to modeling structured reporting is mimicking the interaction of a radiologist with a structured reporting user interface as a series of question-and-answer pairs for a given radiological image. In this direction, we present a hierarchical VQA dataset for structured reporting and a baseline for it in our work Rad-ReStruct [185].

¹<https://radlex.org/>

²<https://www.radelement.org/>

³<https://radreport.org/>

⁴<https://www.befundung.drg.de/>

7.1.2 Evaluating the Clinical Correctness of Reports

Another important aspect of structured reports is that they offer a much more effective and granular evaluation of the clinical correctness of automatically generated radiology reports. This section discusses the implications of this in detail, and while mainly addressing chest X-rays, the conclusions can be transferred to report generation for other medical applications. As we discuss in our work RaDialog [186], automatic radiology report generation from chest X-ray images has been extensively researched. However, evaluating these methods has primarily relied on NLP metrics created for machine translation applications. The problem with these metrics is that while they may be effective in evaluating the reporting style and use of correct medical terms, they are notoriously misleading when assessing clinical correctness. This issue is highlighted by Pino et al. [190], who demonstrated that a simple negation of a sentence leads to similar NLP metrics while obviously changing the clinical implications dramatically. Similarly, Babar et al. [7] showed that a report generation method that does not even consider the given X-ray image can produce competitive NLP metrics.

To remedy this, more recent works have adopted a range of methods to assess the clinical correctness of reports. The most established metric for evaluating the clinical correctness of generated reports in chest X-rays is to extract the CheXpert [108] classification labels from the generated report and compare them with the labels extracted from the ground truth. Commonly, the CheXbert labeler [215] is used for this purpose, and the metric is called the clinical efficacy (CE) score.

To evaluate the correctness of more granular findings than just high-level CheXbert labels, Yu et al. [264] introduced a new metric based on the presence of clinical entities and their relationships in RadGraph, namely RadGraph F1 and RadCliQ. The latter combines RadGraph F1 with the commonly used NLP metrics.

With Rad-ReStruct [185], we have proposed a challenging VQA benchmark that aims to assess the granular image understanding of radiological vision-language models and multimodal LLMs by modeling structured reporting as a VQA task.

Contrastive Language-Image Pre-training for Structured Reporting of Chest X-rays

Contents

8.1	Introduction	75
8.1.1	Related Work	77
8.2	Method	78
8.2.1	Log-Sum-Exp Sign Loss	79
8.2.2	Contrastive Language-Image Pre-training	80
8.2.3	Cross-modal Similarity Metric	80
8.3	Experimental Setup	81
8.3.1	Structured Reporting Dataset	81
8.3.2	Implementation and Training Details	83
8.3.3	Few-shot Classification	85
8.4	Results	86
8.4.1	Ablation and Cardiomegaly Grading	86
8.4.2	Localization of Pathologies	86
8.5	Discussion	88
8.6	Conclusion	89

8.1 Introduction

The process of documentation and report writing often consumes a significant portion of radiologists' time, diverting their attention from addressing individual patient needs [100, 172]. Structured reporting has emerged as a highly valued approach in radiology, offering the potential to streamline this process and standardize the content and terminology of radiological reports. As defined by Nobel et al. [171], structured reporting is an IT-based method that facilitates the import and organization of medical content into a standardized format, enabling the representation of clinical findings in a structured manner.

As discussed in the previous chapter, adopting structured reporting and standardized reports has gained support from prominent radiology societies, including RSNA and ESR. This endorsement stems from the numerous benefits of structured reporting, such as improved communication, enhanced machine readability of reports, and time efficiency. These advantages have far-reaching implications for quality assurance processes, clinical trials, and the

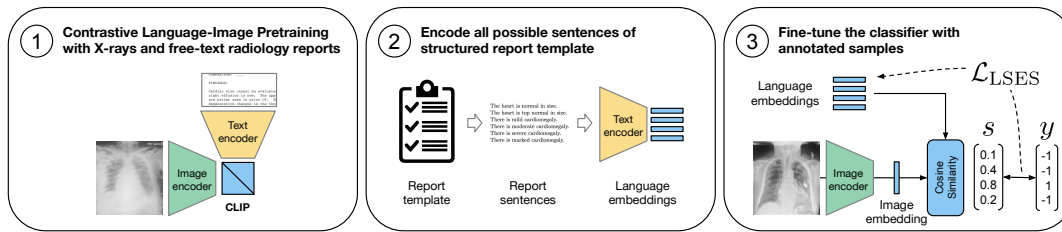


Fig. 8.1. The proposed Few-shot classification with Language Embeddings for chest X-ray Reporting (FlexR) method builds on self-supervised pre-training to accurately predict fine-grained radiological findings, requiring only a limited number of high-quality annotations. The approach consists of three main stages: (1) contrastive language-image pre-training on a dataset containing radiology images paired with their corresponding unstructured reports, (2) encoding of radiological findings extracted from structured reports, and (3) fine-tuning of the resulting language embeddings to optimize the specific structured reporting task. This process enables the efficient extraction of clinical knowledge from unlabelled, unstructured patient data. [122]

internationalization of medical data, as well as the promotion of evidence-based medicine as discussed in Section 2.1.4.

While numerous deep-learning approaches have been proposed for automated reporting, most focus on generating free-text reports rather than structured ones. Although these generated free-text reports offer potential advantages, they may be subject to the same limitations as manually written free-text reports. These include a lack of standardization and challenges in evaluating clinical accuracy [190]. This lack of standardization can also hinder the effective use of written reports as a supervision signal for data-driven decision-support systems.

To address the challenges outlined above, we introduce FlexR [122], an adaptable few-shot learning approach designed to predict fine-grained clinical findings for structured reporting. The proposed method leverages self-supervised pre-training on a large dataset of chest X-rays paired with their corresponding free-text radiology reports. By extracting knowledge from this vast collection of unstructured data, FlexR enables the prediction of structured findings defined by sentences within reporting templates, which can be easily modified to suit specific requirements. Our study's results demonstrate that FlexR can accurately predict the severity of cardiomegaly and localize pathologies in chest X-rays, even when trained with a limited number of image-level annotations.



Contributions:

- We introduce a novel few-shot model for classifying clinical findings to fill structured reports that can easily be adapted to other reporting templates.
- We demonstrate the potential of contrastive language-image pre-training on chest X-rays for structured reporting.
- We evaluate the method on two structured reporting tasks: pathology localization and severity grading of cardiomegaly.

8.1.1 Related Work

The rapid advancements in deep learning and natural language processing (NLP) have paved the way for significant progress in predicting findings and generating radiology reports from X-ray images [101, 243]. This section will focus on the most relevant literature in structured report generation, joint language-image embeddings, and few-shot learning in data-scarce scenarios.

Structured Report Generation

While most radiology report generation research has been focused on the free-text, only a limited number of studies have explored the automation of structured reporting. Pino et al. [190] introduced a structured report generation approach that utilizes a classification model to identify high-level pathologies and select appropriate sentences from a predefined template. This method demonstrates the potential for generating standardized reports that can be easily integrated into clinical workflows while lacking details of a typically written report. Bhalodia et al. [21] employed an object detection algorithm to localize pneumonia and predict additional attributes of the lesions, which could be adapted to a structured reporting setting. Closest to our approach but using full supervision, Syeda-Mahmood et al. [222] classified clinical findings in X-ray images at a fine-grained level using a one-hot-encoded vector, which was then used to retrieve similar reports for the generation of free-text reports.

Joint Language-Image Embeddings

In addition to the CLIP approaches discussed in Section 3.3.3, the generation of vision-language representations has been explored for various tasks in the radiology domain. For instance, Yan et al. [258] investigated weakly-supervised contrastive pre-training for report generation, while Chauhan et al. [43] employed joint embeddings for pulmonary edema assessment. Additionally, Liao et al. [145] utilized mutual information maximization for chest X-ray classification, and Huang et al. [106] proposed an attention-based contrastive learning approach.

Recent studies have also explored various approaches to capture the hierarchical relationships among pathologies in medical imaging. For instance, Pham et al. [189] and Chen et al. [46] proposed methods to model the hierarchical structure of pathologies, enabling a more accurate representation of the complex relationships between different medical conditions. Graph Convolutional Networks (GCNs) have also emerged as a promising technique for leveraging label dependencies in medical image analysis tasks [44, 269]. These GCN-based approaches aim to model pathology dependencies in a graph to improve the performance and interpretability of the models. Furthermore, Zhang et al. [269] introduced a novel method that combines a pre-constructed disease knowledge graph with a report generation module, allowing for the joint learning of visual features and the modeling of relationships between diseases. This approach demonstrates the potential of integrating formal knowledge (see Section 2.1.4) with deep learning.

Few-shot Learning

Few-shot learning has the potential to address the challenge of limited annotated data in medical imaging tasks, particularly in the context of chest X-ray diagnosis. By leveraging the power of few-shot learning, models can learn to make accurate predictions using only a small number of labeled examples. Paul et al. [182] proposed a discriminative autoencoder ensemble that operates in a few-shot setting, demonstrating its effectiveness in diagnosing chest X-rays. Similarly, Jia et al. [114] explored the potential of few-shot learning for generating reports on rare diseases, where obtaining large amounts of labeled data is often impractical.

Recent advancements in zero-shot learning have also shown promise in medical image analysis, particularly using CLIP (Contrastive Language-Image Pre-training) models. These models can make predictions based solely on text embeddings of pathologies without requiring any labeled image data. Seibold et al. [209] and Tiu et al. [230] successfully applied CLIP-based zero-shot classification to chest X-rays using language prompts that describe diseases and their negation. Additionally, Huang et al. [106] and Boecking et al. [25] demonstrated the effectiveness of their improved pre-training techniques in both zero-shot and few-shot settings for predicting chest pathologies.

While these studies have made significant contributions to the field, it is important to note that they primarily focus on classifying the multi-label presence of pathologies or the generation of unstructured reports. In contrast, the work presented here aims to go beyond simple classification by predicting fine-grained labels, such as disease localization and severity grading, which are crucial for providing clinicians with more comprehensive and actionable insights upon which to reason.

8.2 Method

The proposed Few-shot classification with Language Embeddings for chest X-ray Reporting (FlexR) method leverages self-supervised pre-training to predict structured, fine-grained clinical findings from radiology images using text prompts. By leveraging large amounts of unstructured radiology data, FlexR aims to accurately predict radiological findings with few annotated images. The method involves extracting sentences from a structured radiology report template, defining them as potential clinical findings, and projecting them onto a joint language-image embedding space. Using a few annotated samples, these embeddings can be further optimized. At test-time, the fine-tuned language embeddings similar to the encoded image are predicted to fill the structured reporting template. Figure 8.1 illustrates the three key steps of the FlexR method:

- 1. Contrastive language-image pre-training (CLIP)** on a dataset of unlabeled radiology reports and image pairs: we use radiology-specific image and text encoders to initialize our custom CLIP model and pretrain on radiology data.
- 2. Generating language embeddings of clinical findings:** The FlexR method extracts individual sentences from the structured reporting template, representing all possible options, and encodes them using our CLIP text encoder. This process yields a text embedding T_i

for each clinical finding. For example, when detecting and grading cardiomegaly in chest radiographs, the prompts could be formulated as shown in Table 8.1, such as *There is mild cardiomegaly*.

3. Fine-tuning the classifier: The final classifier, which is the output of the FlexR approach, comprises our CLIP image encoder that generates the embedding I_i of the input image and the embeddings $W = T_1, T_2, \dots, T_C$, which are initialized by the text embeddings of the radiological findings and then fine-tuned in the previous step. Here, C denotes the number of clinical findings defined by the structured reporting template. The cosine similarity $s = s_1, s_2, \dots, s_C$ between I_i and each clinical finding in W is computed to classify the input image.

In the final step of the FlexR method, the computed similarities are transformed into predictions by establishing thresholds for mutually non-exclusive findings and selecting the highest similarity for exclusive findings, such as clinical gradings. It is crucial to recognize that, in addition to encoding the presence of clinical pathologies, the FlexR method also encodes the absence of findings as separate embedding, e.g., *The lungs are clear*.

The FlexR method leverages that many textual prompts share common information with other sentences from similar parts of the structured reporting template. This shared information provides a useful clustering of medically similar findings and captures label dependencies. For example, the prompts *lung opacity in the left lung* and *lung opacity in the upper left lung* have nearly identical language embeddings due to their semantic similarity. In rare cases, it was observed that two different prompts might even have the same embedding when using the initialization W . To address this issue and ensure that different prompts result in distinct language embeddings, the FlexR method proposes optimizing the clinical finding embeddings in W and the image encoder using the Log-Sum-Exp Sign loss.

8.2.1 Log-Sum-Exp Sign Loss

To optimize the clinical finding embeddings initialized by the text encoder, we propose using the Log-Sum-Exp Sign (LSES) loss function, as introduced by Jin et al. [116]. The labels of each clinical finding are denoted as $y = y_1, y_2, \dots, y_C$, where $y_i \in \{1, -1\}$, indicating the presence or absence of a clinical finding in the report, respectively. Given the cosine similarity s between the image and finding embeddings, the $\mathcal{L}_{\text{LSES}}$ loss is defined as

$$\mathcal{L}_{\text{LSES}} = \log \left(1 + \sum_{i=1}^C e^{-y_i \gamma s_i} \right). \quad (8.1)$$

The LSES loss inherently assigns higher weights to misclassified classes while leaving the embeddings of correctly initialized classes largely unaltered. The hyperparameter γ allows for the adjustment of this effect, further increasing the loss for misclassified embeddings and decreasing it for well-classified embeddings. This mechanism has proven effective in classification tasks with long-tailed distributions, such as human-object interaction recognition [116], making it well-suited for the long-tailed distribution often encountered in structured

reporting (see Figure 8.2). In the $\mathcal{L}_{\text{LSES}}$ loss, 1 is added to the summands to ensure a lower bound of 0 for the loss.

By employing the LSES loss, the FlexR method can effectively optimize the clinical finding embeddings, prioritizing the correction of misclassified embeddings while preserving the quality of well-initialized embeddings. This approach helps to improve the accuracy and robustness of the predictions, particularly in scenarios with imbalanced class distributions, which are common in structured radiology reporting.

The Log-Sum-Exp (LSE) function is defined as

$$LSE = \log \left(\sum_{i=1}^C e^{x_i} \right), \quad (8.2)$$

serves as a smooth approximation of the maximum function $\max x_1, x_2, \dots, x_i$, with the softmax function being its derivative. By setting $x_i = -y_i \gamma s_i$, the LSE function assigns the highest loss to classes that are either present in the report but have a low similarity with the image embedding or have a high similarity but are not present in the report. Simultaneously, the softmax gradient helps to maintain the stability of the correctly initialized class weights. This behavior makes the LSE function particularly well-suited for optimizing the clinical finding embeddings in the FlexR method.

8.2.2 Contrastive Language-Image Pre-training

FlexR builds on the multimodal embeddings of contrastive language-image pre-training (CLIP) [194], as introduced in Section 3.3.3, that has been demonstrated to be effective for various downstream tasks such as human-object interaction recognition [116]. However, at the time of this work, no models were available that had been trained on a comparable scale for chest radiographs and their corresponding reports. To address this gap, we explored strategies for fine-tuning CLIP on chest radiographs and training a similar model from scratch by initializing both the text and image encoder with domain-specific pretrained models.

8.2.3 Cross-modal Similarity Metric

The FlexR method employs a fully connected layer without bias to compute the cosine similarity between the input image embedding and the language embeddings of the clinical finding prompts. To ensure a well-defined and differentiable similarity metric, both the initial text embeddings and the image embeddings are normalized, and the dot product between these normalized embeddings is calculated using the linear layer. This approach allows the embeddings of the radiological findings to be optimized during training, enhancing the predictions of poorly initialized classes. Algorithm 1 presents the pseudo-code for a PyTorch implementation of the prompt similarity module, defined as a subclass of `nn.Linear` and initialized with the text embeddings of each finding description:

Algorithm 1: Prompt Similarity Pytorch Module

```
class PromptSimilarity(nn.Linear)
  method __init__(prompt_embeddings)
    out_features, in_features ← prompt_embeddings.shape;
    call super().__init__(in_features, out_features, bias=False);
    self.weight.data ← F.normalize(prompt_embeddings);

  method forward(x)
    x ← F.normalize(x);
    x ← super().forward(x);
    return x;
```

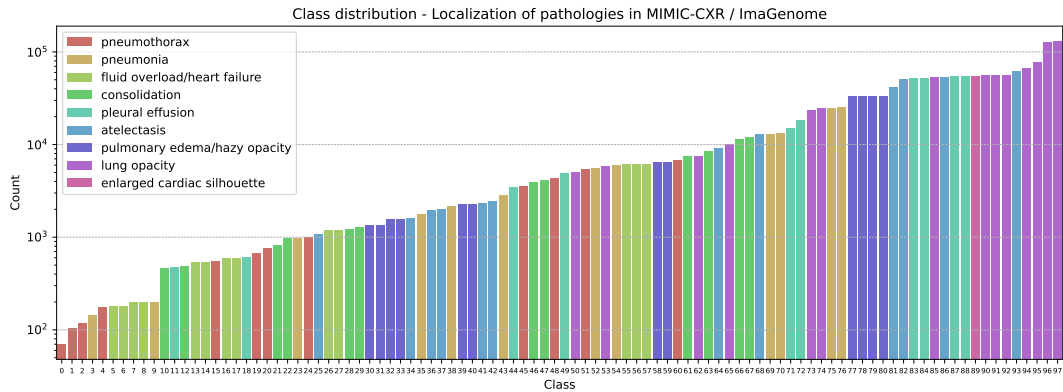


Fig. 8.2. The class distribution in the MIMIC-CXR / ImaGenome task, which involves localizing nine pathologies across 29 anatomical locations, exhibits a long-tailed pattern. The classes, arranged in order of increasing frequency of occurrence and displayed on a logarithmic scale, demonstrate a significant imbalance, with a few classes appearing much more frequently than other classes in the dataset and severely underrepresented classes on the left. [122]

8.3 Experimental Setup

This section provides a detailed description of the experimental setup and the dataset used in our study. We perform domain-specific contrastive pre-training and evaluate the performance of the FlexR method on two structured reporting tasks: assessing the severity of cardiomegaly and localizing pathologies in chest X-rays. Additional information can be found in the appendix.

8.3.1 Structured Reporting Dataset

The dataset employed in this study is the MIMIC-CXR-JPG v2.0.0 [117], which is derived from the MIMIC-CXR dataset. The MIMIC-CXR dataset consists of 377,110 chest radiographs associated with 227,827 imaging studies and free-text reports [81, 118]. To obtain labels for structured reports, we utilize the Chest ImaGenome [253], a medical scene graph dataset containing 242,072 anatomy-centered scene graphs for the MIMIC-CXR image data. The Chest ImaGenome dataset provides 1,256 combinations of relation annotations between 29 anatomical locations and their attributes.

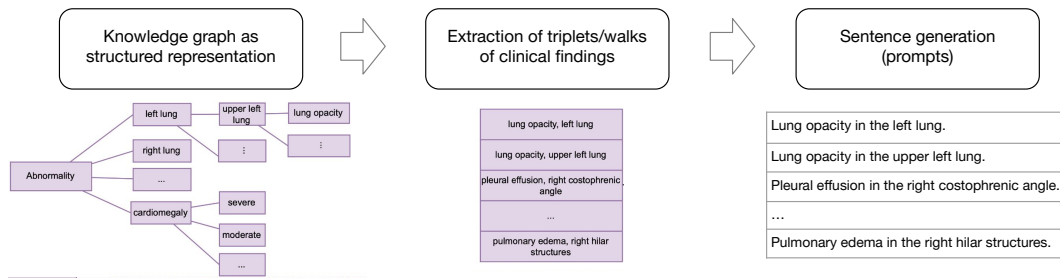


Fig. 8.3. Extraction of structured findings from ImaGenome[253]: Structured reporting elements resembling common data elements (CDEs) are extracted from the ImaGenome knowledge graph by first extracting triplets and then rephrasing them as sentences. [122]

Pathology Localization Reporting Template

Due to the limited availability of publicly accessible reporting templates with annotations, we model the localization of pathologies in chest radiographs as a surrogate reporting task. This task, introduced by AnaXNet [2] and ImaGenome [253], provides annotations for the MIMIC-CXR dataset in the form of graphs. The objective is to detect and localize 9 pathologies: *Lung Opacity*, *Pleural Effusion*, *Atelectasis*, *Enlarged Cardiac Silhouette*, *Pulmonary Edema/Hazy Opacity*, *Pneumothorax*, *Consolidation*, *Fluid Overload/Heart Failure*, and *Pneumonia*.

The dataset encompasses 29 anatomical locations, including various regions of the lung, hilar structures, costophrenic angle, mediastinum, cardiac silhouette, and trachea. We extract the triplet of *pathology* located in the *anatomical site* from the provided graph for each patient. This process, shown in Figure 8.3, results in 98 unique combinations of pathology and location for all patients out of the 162 possible combinations. To create the template sentences used as an initialization of the classifier, we join the *pathology* and *location* with the phrase "in the", for example, "*Consolidation in the left lung*".

Cardiomegaly Severity Reporting Template

For the assessment of cardiomegaly severity, we use the TLAP-endorsed structured reporting template "Chest Xray - 2 Views"¹ illustrated in Figure 8.4. The exact sentences from this template are used as language embeddings, and the associated labels are extracted from the MIMIC-CXR dataset using simple keyword matching. Table 8.1 presents the prompts and their distribution within the dataset.

Data Processing

To train and evaluate the FlexR method, we utilize the data split provided by ImaGenome, which includes both Posterior-Anterior (PA) and Anterior-Posterior (AP) radiographs. After preprocessing, the dataset consists of 166,512 training images, 23,952 validation images, and 47,389 test images. The image processing pipeline is implemented using MONAI 0.8.0. All images are resized to a consistent resolution of 224x224 pixels, with padding applied if necessary, and their pixel values are scaled to the range [-1, 1].

¹created by Penn Medicine: <https://radreport.org/home/144/2011-10-21%2000:00:00>

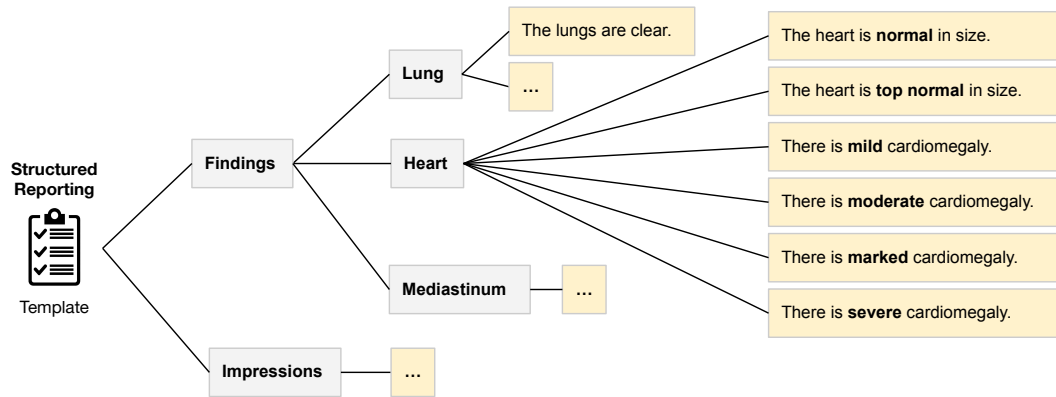


Fig. 8.4. RadReport data elements for severity of cardiomegaly: Structured reports can be represented as a hierarchy of findings often organized by organs of interest.

Several image augmentation techniques are employed during training to improve the model’s robustness and generalization ability. These augmentations include random cropping, ensuring that at least 75% of the original image size is retained; random rotation, allowing for angles up to $\pm 15^\circ$; color jittering, with a brightness variation of 10%; and contrast and saturation variations of 20%.

The corresponding reports are augmented by randomly sampling a sentence containing a finding from the ImaGenome scene graph. For healthy patients with no findings, a random sentence from the full report is sampled instead. This augmentation strategy helps to expose the model to a diverse range of clinical findings and their associated textual descriptions, enhancing its ability to learn meaningful representations and make accurate predictions.

8.3.2 Implementation and Training Details

All networks in this study are trained using PyTorch 1.10 and PyTorch Lightning 1.5.10 in native mixed precision. The transformer-based models and tokenizers are implemented using the Hugging Face library (transformers 4.16.2). A single NVIDIA A40 GPU is used for training all classification models. The DenseNet and Vision Transformer classifiers are trained for 25 epochs, following the hyperparameters used in [2]: an Adam optimizer with a learning rate of $1e-4$ and an unweighted binary cross-entropy loss.

The FlexR models are fine-tuned for 10 epochs using a learning rate of $1e-4$, an AdamW optimizer without weight decay, and a learning rate scheduler with cosine annealing decay and a 1-epoch linear warmup. During the fine-tuning process, all weights, including those of the image encoder, are optimized. Experiments involving fine-tuning only the language embeddings and zero-shot inference did not yield useful results and were therefore excluded from the study. The model with the best performance on the validation set is selected for testing. After a hyperparameter search on the localized pathology task with values of 50, 100, and 150, the γ parameter of the LSES loss is set to 50 for all experiments.

An epoch exposes the model to 128 images per class in the context of few-shot learning. The same number of images is chosen per class during sampling to ensure comparability while

accessing the entire dataset. A batch size of 256 is used for the classification baselines and FlexR fine-tuning. All experiments are repeated 5 times with different random seeds, and the average results are reported.

The DenseNet121 baseline is trained using a cross-entropy loss for the cardiomegaly severity assessment task. To enable a fair comparison with our custom CLIP model, the DenseNet121 used as a baseline in the two few-shot tasks is also initialized with a model pre-trained on detecting pathologies in MIMIC-CXR without localization.

Pre-training details

We utilize all reports and images in the MIMIC-CXR training set for contrastive pre-training. Initial experiments involving fine-tuning the original CLIP model with a ViT-16/B backbone [65] using the original CLIP tokenizer and text encoder resulted in severe overfitting. To address this issue, we replace the image and text encoders with domain-specific pre-trained encoders: DenseNet121 (DN121) [102] pre-trained on the detection of pathologies and SciBERT (SB) [17] pre-trained on a large corpus of scientific text. Contrastive pre-training with these encoders demonstrates better generalization, and consequently, we employ this model for all subsequent experiments, comparing it with the initially fine-tuned, unmodified ViT-B/16-CLIP model. Additional pre-training details are provided in the appendix.

The CLIP models are fine-tuned on all radiology reports and chest X-ray images in the MIMIC-CXR training dataset using 8 NVIDIA A40 GPUs for 300 epochs with a batch size of 128. An AdamW optimizer is used with a learning rate 5e-6, no weight decay on the normalization layer and bias, and a weight decay of 0.1 on all other parameters. The learning rate is decayed using a single cosine annealing schedule with a 1-epoch linear warmup. The Hugging Face library (transformers 4.16.2) implements the vision-language models.

Initially, we fine-tuned the weights provided by OpenAI for the ViT-B/16-CLIP configuration². However, upon observing overfitting on the MIMIC-CXR dataset, we replace the image encoder with a DenseNet121 and the language encoder with SciBERT³. The DenseNet121 is pre-trained on the detection of pathologies without localization (see Section 8.3.3) on the MIMIC-CXR training set. As the embedding dimensions of the image and text encoders may differ, their embeddings are projected to a joint embedding size of 512 using a linear layer.

Following the original CLIP paper [194] and its implementation in Hugging Face, we calculate the cosine similarity between the text and image embeddings and apply a contrastive loss to encourage similarity between corresponding pairs of text and image inputs while pushing apart non-corresponding pairs. Specifically, the contrastive loss is a cross-entropy loss applied to both the rows and columns of the similarity matrix, and the average of these two losses is used for backpropagation.

²<https://huggingface.co/openai/clip-vit-base-patch16>

³https://huggingface.co/allenai/scibert_scivocab_uncased

Tab. 8.1. Grading for cardiomegaly severity in the *Chest Xray - 2 Views* RadReport template with the label support extracted from MIMIC-CXR reports. [122]

Severity	Initialization prompt	Training	Validation	Testing
Normal	The heart is normal in size.	3140	478	943
Top Normal	The heart is top normal in size.	635	72	160
Mild	There is mild cardiomegaly.	6084	809	1816
Moderate	There is moderate cardiomegaly.	8696	1164	2619
Severe	There is severe cardiomegaly.	2231	335	676
Marked	There is marked cardiomegaly.	246	36	85
		21032	2894	6299

8.3.3 Few-shot Classification

This study’s leading set of experiments focuses on evaluating the effectiveness of the FlexR method in utilizing knowledge extracted through contrastive pre-training for the few-shot classification of fine-grained clinical findings. We deliberately reduce the training data and evaluate the models in an N-shot setting to assess the model’s performance in scenarios with limited annotated data. N-shot refers to using N annotated samples per class. Furthermore, to establish an upper bound for the model’s performance, we compare the results obtained in the few-shot settings with those achieved using all available annotated data. This comparison provides valuable insights into the potential performance gap between few-shot learning and fully supervised learning, as well as the effectiveness of the FlexR method in bridging this gap by leveraging knowledge acquired through contrastive pre-training.

Severity of Cardiomegaly

The first few-shot experiment evaluates the model’s ability to adapt to a new structured reporting workflow defined by an existing reporting template. The performance of the FlexR method and other baseline models is evaluated using the area under the receiver operating characteristic curve (AUC). This metric comprehensively assesses the model’s ability to discriminate between different cardiomegaly severity levels across various decision thresholds. Using an existing structured reporting template and extracting labels from the MIMIC-CXR dataset, this experiment aims to simulate a realistic scenario where the FlexR method is applied to a new structured reporting workflow. Taking exact sentences from the template as language embeddings ensures the model is evaluated on clinically relevant and meaningful prompts.

Pathology Localization

Following the approach of Agu et al. [2], we calculate the area under the receiver operating characteristic curve (AUC) for all possible locations of each pathology and average them to obtain a single location-sensitive AUC per pathology. This evaluation metric comprehensively assesses the model’s ability to detect and localize pathologies across anatomical sites. The pathology localization task is employed in the ablation study as it offers more available annotations than other tasks. By modeling the localization of pathologies as a surrogate reporting

task, we can evaluate the performance of the FlexR method and other baseline models in a clinically relevant context, even without extensively annotated reporting templates.

8.4 Results

8.4.1 Ablation and Cardiomegaly Grading

The ablation study, summarized in Table 8.2, focuses on the task of localizing pathologies and demonstrates that the FlexR model outperforms both the random initialization without language embeddings and the ViT-B/16-CLIP backbone fine-tuned on the MIMIC-CXR dataset. These results confirm the superior generalization capability of the domain-adapted CLIP model and highlight the importance of initializing the model with language embeddings for improved few-shot performance. Furthermore, FlexR surpasses the performance of a naïve transfer learning baseline using a DenseNet121 pre-trained on detecting non-localized pathologies. As the number of samples per class increases, the performance gap between FlexR and the other models diminishes.

The second part of Table 8.2 presents the results of detecting and grading cardiomegaly severity using language embeddings extracted from a real-world RadReport reporting template. In the 1-shot learning scenario, FlexR increases the area under the receiver operating characteristic curve (AUC) of 0.06 compared to the naïve transfer learning baseline. This improvement grows to 0.07 in the 5-shot learning setting. Interestingly, the DenseNet121 model optimized using cross-entropy loss required oversampling of underrepresented classes to achieve a performance similar to that of FlexR. In contrast, FlexR reached its best AUC of 0.82 without oversampling when trained on all available data. This observation could be attributed to the inherent class weighting property of the LSES loss function employed by FlexR.

These findings underscore the effectiveness of the FlexR method in adapting to new structured reporting workflows and accurately grading the severity of cardiomegaly, even in few-shot learning scenarios. Furthermore, the ablation study shows the effectiveness of the domain-adapted CLIP model and the initialization with language embeddings before fine-tuning.

8.4.2 Localization of Pathologies

FlexR performs better than transfer learning with a pre-trained DenseNet121 in the few-shot setting of localized pathology detection. Specifically, FlexR achieves a 0.07 higher area under the receiver operating characteristic curve (AUC) for 1-shot learning and a 0.08 increase for 5-shot learning. Table 8.3 compares the performance of FlexR with global classification baselines and object detection-based methods that utilize the full training data. To establish an upper bound for the AUC, we report the performance of image encoders used for global pathology detection without localizing the diseases. As expected, the few-shot methods do not reach the AUC of AnaXNet, which leverages all available training data, is fully supervised with bounding box annotations, and refines features extracted from high-resolution image crops.

Tab. 8.2. Ablation study of the FlexR method using different backbones and weight initializations for pathology localization, along with results for cardiomegaly grading compared to naïve transfer learning. The evaluation metric is mean AUC. N-shot indicates the number of annotated samples per class used for training. The proposed approach is highlighted in **bold**. [122]

Method	Backbone	Pretraining	1-shot	5-shot	10-shot	100-shot	sampled	all
<i>Ablation on localizing pathologies</i>								
MLP	DN121	pathologies	0.67	0.69	0.71	0.76	0.77	0.84
FlexR	ViT-B/16-CLIP	CLIP	0.66	0.70	0.73	0.75	0.77	-
FlexR	DN121+SB	random init.	0.67	0.72	0.75	0.79	0.81	-
FlexR	DN121+SB	CLIP	0.74	0.77	0.78	0.80	0.81	0.84
<i>Grading task: Cardiomegaly severity prediction</i>								
MLP	DenseNet121	pathologies	0.59	0.65	0.68	0.75	0.79	0.73
FlexR	DN121+SB	CLIP	0.65	0.72	0.74	0.77	0.78	0.82

Tab. 8.3. Comparison of FlexR against baselines using all available data, with and without pathology localization, as well as naïve transfer learning in the few-shot setting. The evaluation metric is AUC, and N-shot refers to the number of annotated samples per class used for training. The proposed method is marked in **bold** [122].

Method	Lung Opac.	Pleural Eff.	Atelectasis	Enl. Card. S.	Pulm. Edema	Pneumothor.	Consolidation	Heart Failure	Pneumonia	Avg. AUC
<i>Multi-label classification with no localization on global view using all data</i>										
DenseNet169 [2]	0.91	0.94	0.86	0.92	0.92	0.93	0.86	0.87	0.84	0.89
DenseNet169	0.87	0.90	0.79	0.86	0.85	0.83	0.75	0.77	0.75	0.82
DenseNet121	0.88	0.91	0.81	0.87	0.87	0.87	0.79	0.80	0.77	0.84
ViT-B16	0.88	0.91	0.80	0.87	0.86	0.85	0.77	0.78	0.76	0.83
<i>Fully supervised object detection with bounding boxes and high-resolution crops using all data</i>										
FasterR-CNN [2]	0.84	0.89	0.77	0.85	0.87	0.77	0.75	0.81	0.71	0.80
AnaXNet [2]	0.88	0.96	0.92	0.99	0.95	0.80	0.89	0.98	0.97	0.93
<i>Few-shot, detector-free localization on global view (224 × 224)</i>										
DenseNet121 1-shot	0.70	0.76	0.64	0.77	0.70	0.60	0.66	0.62	0.58	0.67
DenseNet121 5-shot	0.72	0.78	0.66	0.78	0.73	0.64	0.67	0.64	0.62	0.69
DenseNet121 (all data)	0.83	0.89	0.79	0.87	0.84	0.89	0.83	0.81	0.82	0.84
FlexR 1-shot	0.72	0.83	0.69	0.82	0.77	0.72	0.74	0.73	0.67	0.74
FlexR 5-shot	0.75	0.84	0.71	0.82	0.79	0.78	0.76	0.73	0.71	0.77
FlexR (all data)	0.82	0.89	0.78	0.87	0.84	0.90	0.83	0.80	0.81	0.84

Still, the results highlight the effectiveness of the FlexR method in detecting and localizing pathologies in few-shot learning scenarios. Despite the limited training data, FlexR outperforms the transfer learning approach with a pre-trained DenseNet121, demonstrating its ability to adapt efficiently to new pathologies and anatomical locations. However, it is important to acknowledge that the few-shot methods, including FlexR, do not achieve the same level of performance as fully supervised object detection methods like AnaXNet, which benefit from the supervision signal of bounding box annotations for each pathology. Nevertheless, the superior performance of FlexR in the few-shot setting underscores its potential for flexible adaptation to new structured reporting tasks and its ability to provide accurate pathology detection and localization with limited annotated data.

8.5 Discussion

Modeling radiology report generation as a classification task offers several advantages, such as allowing for a direct evaluation of the clinical correctness of reports and aligning with the growing adaptation of structured reporting and standardized reports. However, capturing the nuances of radiology reports requires a highly fine-grained classification of clinical findings. While unstructured data, such as free-text reports, are abundant, detailed, structured, and high-quality annotations are scarce. To address this challenge, we have introduced a method that leverages unstructured data to learn and predict fine-grained clinical findings using only a few annotated samples per class. Our results demonstrate that our method can be easily adapted to hospital-specific reporting templates and outperforms the baseline in the severity assessment of cardiomegaly and the localization of pathologies in chest X-rays in a few-shot learning setting.

It is important to note that self-supervised pre-training with CLIP and FlexR is only feasible when large amounts of domain-specific image-text pairs are available, which may not be true in all medical applications. Additionally, our results indicate that using all available labels for training outperforms few-shot learning approaches. Therefore, if possible, all labels should be utilized during training. Specifically, our findings suggest that the localization of diseases in chest X-rays is better suited for a specialized object-detection-based model that uses full supervision with bounding boxes. Unlike the approaches proposed by Seibold et al. [209] and Tiu et al. [230], our method employs only a single negative prompt, representing a healthy patient without any findings, rather than negative prompts for every pathology. Adapting this strategy to include two reference embeddings could potentially improve the performance in detecting diseases.

Moreover, label dependencies (e.g., located in the *left lung* and *lower left lung*) have been modeled implicitly in FlexR through the similarity of text in prompts. These dependencies could be explicitly modeled in the future to further enhance the method's performance. Ultimately, the greatest potential for improvement lies in refining the pre-training of joint vision-language representations. This objective can be achieved by utilizing additional data or developing improved methodologies to address the severe class imbalance of clinical findings.

A publicly available, standardized reporting template for chest X-rays, featuring a high level of detail and corresponding annotations for datasets like MIMIC-CXR, is essential to facilitate a more comprehensive evaluation of structured reporting in the future. While Wu et al. [253] and Jain et al. [110] provide highly detailed, structured annotations in the form of graphs, there is a need for a translation of these annotations to a real-life reporting template. Establishing a benchmark for comprehensive, structured reporting would facilitate future research in this direction, enabling the development and comparison of advanced automated radiology report generation methods.

8.6 Conclusion

In this work, we emphasize the importance of developing methods for structured reporting in radiology, as standardized reports can formalize the evaluation of knowledge embedded in the representations of neural networks. To address this need, we propose the Few-shot classification with Language Embeddings for chest X-ray Reporting (FlexR) method, which leverages self-supervised pre-training with CLIP to predict fine-grained clinical findings for a given radiology image.

Our results demonstrate that, even with limited image-level annotations, the FlexR method can effectively predict the structured reporting subtasks of cardiomegaly severity assessment and localizing pathologies in chest X-rays in a few-shot learning setting. These findings highlight the potential of our approach to enable the generation of detailed, standardized radiology reports with minimal annotated data, thereby facilitating the adoption of structured reporting in clinical practice.

The extraction of fine-grained radiological findings from images instead of directly predicting a disease with a deep learning model is important. It allows us to intuitively verify the correct image understanding of a model. In addition to facilitating structured reporting, these detailed findings can also be integrated with other non-imaging patient information in clinical reasoning support systems. With the advancement of multimodal, self-supervised pre-training methods and the ability of LLMs to extract structured information from unstructured text, we will be able to model the structured reporting task on a large scale in the future.

Zero-shot Classification of Chest X-rays with Deductive Reasoning on Radiological Findings

Contents

9.1	Introduction	91
9.2	Methodology	92
9.2.1	Model Overview	92
9.2.2	Prompt Engineering	94
9.3	Experiments and Results	94
9.3.1	Ablation Studies	96
9.3.2	Qualitative Results	99
9.4	Discussion	100
9.5	Conclusion	101

9.1 Introduction

Computer-aided diagnosis (CAD) systems have emerged as valuable tools in medical diagnosis, but their effectiveness is constrained by the requirement for extensive labeled data for their training. This limits their adaptation to clinical applications with limited data like rare and emerging diseases [75, 192]. Furthermore, it does not allow for flexible integration in new clinical environments, such as adapting to new reporting templates or to changed guidelines with new definitions of radiological findings. To address these challenges, recent research has explored zero-shot [25, 106, 209, 230, 244] and few-shot [25, 106, 122] learning techniques. These approaches leverage contrastive pre-training [194, 271] on paired radiology reports and images, demonstrating performance comparable to radiologists [230]. However, the black-box nature of these models and the lack of detailed findings limit their interpretability and application. This is important since providing diagnostic explanations using radiological findings can be essential for building trust in the system and enabling radiologists to validate the results [156] as motivated in Section 2.2.3.

Drawing inspiration from the successful application of large language models (LLMs) in predicting image descriptors for natural images [160], we present Xplainer [184]. This method adapts a classification-by-description method of vision-language models to the multi-label setting of classifying radiological findings in medical images. In this approach, the model is tasked with classifying the presence of descriptive observations that a radiologist would look for in a radiograph to confirm a suspected diagnosis. This design choice enables the model

to form a set of radiological findings that must be present in the image to confirm or reject the hypothesis of a suspected disease, testing them against the given image using similarity measures. Xplainer’s approach resembles the deductive, analytical aspect of clinical reasoning discussed in Section 2.1.3 by breaking down the diagnosis into more interpretable image observations that can be verified and documented. This model provides inherent interpretability, as the final prediction is based on the probabilities of the underlying image observations, providing a clear and transparent decision-making process. The prior knowledge in the form of textual descriptions is first generated by an LLM and then refined by a radiologist.

To assess the effectiveness of Xplainer, we conduct evaluations on two well-established chest X-ray datasets: CheXpert [108] and ChestX-ray14 [242]. Our results in these datasets demonstrate that Xplainer offers better performance while providing detailed decision-making insights. In summary, Xplainer introduces a novel approach to zero-shot classification in radiology that enhances both interpretability and diagnostic performance.



Contributions:

- We introduce Xplainer, a novel framework for explainable zero-shot diagnosis from X-ray images, adapting a classification-by-description approach to mirror the analytical reasoning process of radiologists.
- We integrate formal knowledge about the diseases and their manifestations in the form of radiological text descriptions in the model.
- Our work demonstrates that classifying descriptive observations, instead of directly predicting a diagnosis, improves performance and provides intrinsic explainability.
- We show that Xplainer outperforms previous zero-shot methods on the CheXpert and ChestX-ray14 datasets.

9.2 Methodology

9.2.1 Model Overview

We introduce Xplainer, an intrinsically interpretable zero-shot approach for diagnosing pathologies from X-ray images employing a classification-by-description approach. The objective is the multi-label classification of radiological findings in an image i given a set of clinical observations o_{p_1-n} per pathological finding p .

Our zero-shot approach builds upon BioVil [25], a pretrained vision-language model for radiology as described in Section 3.3.3. Using the language and vision encoders from BioVil, we compute the cosine similarity between an input image i and each of N pre-defined clinical observations o_{p_1-N} describing a pathology. By normalizing this similarity to a range from 0

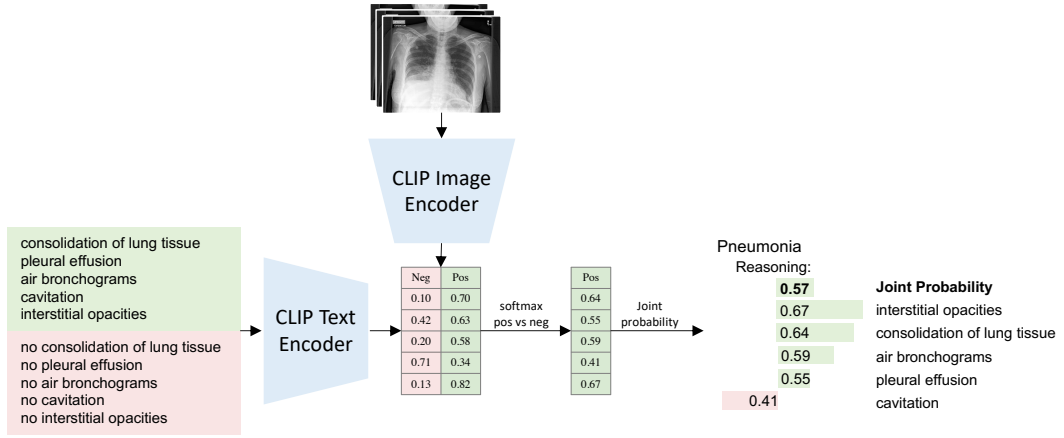


Fig. 9.1. Xplainer overview: Observation probabilities are calculated using contrastive CLIP prompting and then utilized to make an explainable diagnosis prediction, providing insights into the model’s decision-making process. [184] *Reproduced with permission from Springer Nature.*

to 1, we calculate the observation probabilities $P_{pos}(o_i)$ for every observation. Similarly, we calculate probabilities for the absence of all observations $P_{neg}(o_i)$ by defining negated prompts for each observation. The final probability of an observation $P(o_i)$ is determined using the softmax over the positive and negative probabilities. Based on these positive probabilities, we calculate a joint probability to determine the probability of the presence of a pathology $P(p)$:

$$\log(P(p)) = \sum_{i=1}^N \log(P(o_i)) \div N \quad (9.1)$$

The process of calculating the log probability of the presence of a pathology is repeated for all pathologies we aim to diagnose in the image. Our method is inherently explainable by directly extracting the pathology diagnosis prediction from the observation probabilities, producing both a diagnosis prediction and the estimated X-ray observations that led to that prediction. Furthermore, the observation probabilities provide insights into which observations the model primarily considers for its diagnosis, enhancing the interpretability of the decision-making process. Figure 9.1 presents an overview of our framework, illustrating the steps involved in the explainable zero-shot classification-by-description approach for diagnosing pathologies from X-ray scans.

To incorporate multiple images of a single patient, we calculate each image’s positive and negative observation probabilities and then average these probabilities before computing the final pathology probability. This approach allows the model to consider information from multiple views or time points, providing a more comprehensive assessment of the patient’s condition.

9.2.2 Prompt Engineering

The success of zero-shot classification depends heavily on the alignment between the contrastive pre-training and the task at hand [194]. Since BioVil [25] was trained on pairs of X-ray images and reports, it is crucial to phrase prompts in a way that closely resembles the style of radiology reports. To generate our descriptive prompts, we utilize ChatGPT [42], querying it to describe observations found in X-rays that could appear in a radiological report indicating a pathology. These prompts are then refined with the assistance of a radiologist who verifies and adapts the observation descriptors. The final lists of descriptors per chest X-ray class are presented in Table 9.1.

Radiology reports often include both the presence and absence of specific observations. When comparing a prompt with an image embedding, the model may struggle to differentiate between an observation's positive and negative occurrence due to their similar formulation. Previous research [209, 230] has demonstrated that introducing contrastive prompts can address this issue. By prompting the model with both a positive and a negated formulation, we don't have to define a threshold. Instead, we can apply a softmax over them. We further adapt our prompts in two additional steps to align our prompts with the text in radiology reports. First, we include a disease indication, as radiology reports typically contain observations and conclusions. This reduces the ambiguity since a single observation (e.g., lung opacity) can be indicative of multiple pathologies (e.g., pneumonia, atelectasis, or edema). Additionally, we phrase all our observations in a style that resembles an actual report by adding "There is/are" before each observation. The resulting prompt structure for all observations is as follows: "There is/are (no) <observation> indicating <pathology>."

Finally, we compare our observation-based prompting with a baseline using contrastive pathology-based prompts. Similar to previous works, only two prompts are used in this setting: one positive and one negative for each pathology. To demonstrate the benefit of observation-based, contrastive prompting with disease indication and report style, we compare the following styles of prompting:

- **Pathology-based:** (No) <pathology>
- **Basic:** Only positive prompt per pathology: <observation>
- **Contrastive:** (No) <observation>
- **Pathology Indication:** (No) <observation> indicating <pathology>
- **Report Style:** There is/are (no) <observation> indicating <pathology>

9.3 Experiments and Results

We evaluate Xplainer in a zero-shot setting on two widely used chest X-ray datasets: CheXpert [108] and ChestX-ray14 [242]. CheXpert provides a validation set with 200 patients and a test set with 500 patients, encompassing 14 classes, including "No Finding", "Support Devices", and 12 pathology labels. ChestX-ray14 is evaluated on 14 pathology labels using a test set of 25,596 images. For both datasets, we perform multi-label classification and assess the

Tab. 9.1. Descriptors for each pathology used in the Xplainer framework, showcasing the detailed radiological findings that contribute to the diagnosis prediction and explainability of the model. [184] *Reproduced with permission from Springer Nature.*

enlarged cardiome-diastinum	increased width of the heart shadow, widened mediastinum, abnormal contour of the heart border, fluid or air within the pericardium, mass within the mediastinum
cardiomegaly	increased size of the heart shadow, enlargement of the heart silhouette, increased diameter of the heart border, increased cardiothoracic ratio, displaced or elevated diaphragm
lung opacity	increased density in the lung field, whitish or grayish area in the lung field, obscured or blurred margins of the lung field, loss of normal lung markings within the opacity, air bronchograms within the opacity, fluid levels within the opacity, silhouette sign loss with adjacent structures
lung lesion	consolidation of lung tissue, pleural effusion, cavities or abscesses in the lung, abnormal opacity or shadow in the lung, irregular or blurred margins of the lung
edema	blurry vascular markings in the lungs, enlarged heart, kerley b lines, increased interstitial markings in the lungs, widening of interstitial spaces
consolidation	loss of lung volume, increased density of lung tissue, obliteration of the diaphragmatic silhouette, presence of opacities, blunting or loss of sharpness of costophrenic angles
pneumonia	consolidation of lung tissue, pleural effusion, air bronchograms, cavitation, interstitial opacities
atelectasis	increased opacity, volume loss of the affected lung region, displacement of the diaphragm, blunting of the costophrenic angle, shifting of the mediastinum
pneumothorax	tracheal deviation, deep sulcus sign, increased radiolucency, flattening of the hemidiaphragm, absence of lung markings, shifting of the mediastinum
pleural effusion	blunting of costophrenic angles, opacity in the lower lung fields, mediastinal shift, reduced lung volume, presence of meniscus sign or veil-like appearance
pleural other	pleural thickening, pleural calcification, pleural masses or nodules, pleural empyema, pleural fibrosis, pleural adhesions
fracture	visible breaks in the continuity of the bone, misalignments of bone fragments, widening or narrowing of the bone, displacements of bone fragments, disruptions of the cortex or outer layer of the bone, visible callus or healing tissue, fracture lines that are jagged or irregular in shape, multiple fracture lines that intersect at different angles
support devices / foreign objects	artificial joints or implants, stents or other vascular devices, prosthetic devices or limbs, breast implants, radiotherapy markers or seeds
infiltration	irregular or fuzzy borders around white areas, blurring, hazy or cloudy areas, increased density or opacity of lung tissue, air bronchograms
mass	calcifications or mineralizations, dark areas or voids in the scan, shadowing, distortion or compression of tissues, anomalous structure or irregularity in shape
nodule	nodular shape that protrudes into a cavity or airway, distinct edges or borders, calcifications or speckled areas, small round oral shaped spots, white shadows
emphysema	flattened hemidiaphragm, pulmonary bullae, hyperlucent lungs, horizontalisation of ribs, barrel chest
fibrosis	reticular shadowing of the lung peripheries, volume loss, thickened and irregular interstitial markings, bronchial dilation, shaggy heart borders
pleural thickening	thickened pleural line, loss of sharpness of the mediastinal border, calcifications on the pleura, lobulated peripheral shadowing, loss of lung volume
hernia	bulge or swelling in the abdominal wall, protrusion of intestine or other abdominal tissue, swelling or enlargement of the herniated sac or surrounding tissues, retro-cardiac air-fluid level, thickening of intestinal folds

performance using the Area Under the ROC curve (AUC) between the positive pathology probabilities and the corresponding labels.

Tab. 9.2. AUC scores for zero-shot pathology classification on CheXpert and ChestX-ray14 datasets, comparing different prompting approaches. Results marked with * indicate in-domain testing, as the underlying CLIP model was also trained on the ChestX-ray14 dataset, while the other results report out-of-domain performance. [184] *Reproduced with permission from Springer Nature.*

	CLIP pre-training data	CheXpert		ChestX-ray14
		val	test	test
CheXzero [230]	MIMIC	N/A	74.73	-
Seibold et al. [209]	MIMIC	78.86	N/A	71.23
Seibold et al. [209]	MIMIC, PadChest, ChestX-ray14	83.24	N/A	78.33*
Xplainer	MIMIC	84.92	80.58	71.73

Table 9.2 presents our results compared to previously proposed zero-shot pathology prediction approaches. For the CheXpert dataset, we compare our performance with Seibold et al. [209] on the validation set, as they only reported validation performance. When comparing with CheXzero [230] and evaluating on the ChestX-ray14 dataset, we use the test set results. Our approach outperforms both previous works in an out-of-domain setting, where the zero-shot inference is performed on a dataset different from the one used to train the underlying CLIP model. These state-of-the-art results on both datasets demonstrate the effectiveness of our observation-based modeling approach, which aligns with the importance of incorporating clinical knowledge and reasoning into clinical decision support, as discussed in Section 2.2. Table 9.3 provides a detailed breakdown of our results per pathology and dataset, offering further insights into the performance of our method.

9.3.1 Ablation Studies

Our ablation studies examine the effect of prompt design and the impact of using multiple views on the performance of Xplainer. Table 9.4 presents the results on the CheXpert validation set using different prompting styles. We find that pathology-based prompting, which achieves an AUC of 76.14%, performs considerably worse than observation-based prompting, which shows an AUC of 84.92%. This result further emphasizes the benefit of observation-based prompting, aligning with the importance of incorporating detailed clinical knowledge in the reasoning process as discussed in Section 2.1.4. When comparing basic observation-based prompting, which uses only positive prompts per observation, to contrastive prompting, we observe a substantial performance gap. This difference highlights the importance of using negative prompts to differentiate between positive and negative occurrences of observations for effective decision-making. This is reminiscent of the semantic qualifiers used by experienced clinicians described in Section 2.1.3. We also demonstrate the impact of formulating our prompts with pathology indication and in report style. Adding pathology indication to contrastive observation-based prompting improves performance, reaching an AUC of 84.35%. Lastly, incorporating report style in the prompts leads to the highest AUC of 84.92%, suggesting that a contrastive observation-based prompt with pathology indication and report style is the most effective approach for zero-shot prediction of findings in X-ray.

Tab. 9.3. AUC per chest X-ray class of clinical findings on the CheXpert validation and test set as well as the ChestX-ray14 test set. [184] *Reproduced with permission from Springer Nature.*

	CheXpert Val	CheXpert Test	ChestX-ray14
No Finding	88.82	89.94	-
Enlarged Cardiome-diastinum	79.23	80.60	-
Cardiomegaly	78.62	83.32	79.71
Lung Opacity	88.18	91.76	-
Lung Lesion	91.46	69.33	-
Edema	84.84	84.55	81.46
Consolidation	91.56	85.89	71.87
Pneumonia	85.68	83.73	70.83
Atelectasis	84.64	85.46	66.86
Pneumothorax	78.09	83.75	72.18
Pleural Effusion	88.72	89.30	79.11
Pleural Other	83.92	58.67	-
Fracture	-	60.47	-
Infiltration	-	-	68.81
Mass	-	-	70.28
Nodule	-	-	64.74
Emphysema	-	-	74.02
Fibrosis	-	-	62.25
Pleural Thickening	-	-	67.44
Hernia	-	-	74.60
Support Devices / Foreign Objects	80.25	81.15	-

Furthermore, we compare the performance of prompts directly generated by ChatGPT with our expert-refined prompts, as shown in Table 9.5. The refinement process involved removing irrelevant, redundant, or incorrect descriptors from the ChatGPT-generated prompts. We observed an improvement in performance after refining the prompts, indicating that incorporating domain knowledge can further enhance the effectiveness of our method. This aligns with the importance of integrating clinical expertise and knowledge into clinical decision support systems, as discussed in Section 2.2. However, it is worth noting that even the unrefined ChatGPT prompts perform remarkably, demonstrating the potential of combining the knowledge embedded in large, generic language models with specialized domain-specific vision-language models. This highlights the value of leveraging recent NLP advances for incorporating textual knowledge and multimodal representation learning to develop more accurate and explainable models for clinical applications (see Chapter 3).

For the "No Finding" class, we compare two approaches: (1) defining specific prompts such as "Clear lung fields" or "Normal heart size and shape" to classify "No Finding," (prompting) and

Tab. 9.4. Comparison of different prompting styles on the CheXpert validation set, demonstrating the effectiveness of contrastive observation-based prompting with pathology indication and report style. [184] *Reproduced with permission from Springer Nature.*

	AUC
Contrastive pathology-based Prompting	76.14
Observation-based Prompting:	
Basic Prompt	58.65
Contrastive Prompt	77.00
+ pathology Indication	84.35
+ Report Style	84.92

Tab. 9.5. Comparison of ChatGPT-generated prompts and prompts refined with the help of a senior radiologist, showing the benefit of incorporating domain knowledge into prompt engineering. [184] *Reproduced with permission from Springer Nature.*

	CheXpert Val	CheXpert Test	ChestX-ray14
ChatGPT prompts	83.61	79.94	71.40
Refined Prompts	84.92	80.58	71.73

(2) modeling it as the absence of all the other 13 labels (rule-based). As shown in Table 9.6, a rule-based modeling of this class yields better results. This observation suggests that there may not be a clearly defined set of observations that a radiologist would consistently mention in their report to indicate a healthy X-ray scan. The absence of abnormal findings, rather than the presence of specific normal observations, appears to be a more reliable indicator of the "No Finding" class. This approach aligns with the clinical reasoning process of ruling out hypotheses about pathologies in the differential, based on the presence or absence of radiological signs, as discussed in Section 2.1.3.

In the final part of our ablation study, we explore the impact of using a single frontal view versus different aggregation methods of all available views on pathology detection. The aggregation process begins by calculating the positive and negative observation probabilities for each image. In the Maximum aggregation approach, the highest observation probability is selected. The rationale behind this method is that certain perspectives may provide a clearer view of an observation, and the model's most confident perspective should be prioritized. However, integrating multiple views equally can offer complementary insights into image

Tab. 9.6. Comparison of modeling the "No Finding" label using explicit prompts or a rule-based definition as the absence of other findings, demonstrating the effectiveness of the rule-based approach. [184] *Reproduced with permission from Springer Nature.*

	AUC - No Finding
Explicit Prompting	79.64
Rule-based	88.82

Tab. 9.7. Comparison of single-view inference and different methods for multi-view aggregation, highlighting the advantage of averaging observation probabilities across multiple views. [184] *Reproduced with permission from Springer Nature.*

	mean AUC
Frontal view	84.19
Maximum	84.77
Mean	84.92

observations. To capitalize on this type of multi-view information, we evaluate the Mean aggregation technique, where observation probabilities are averaged across multiple images. Table 9.7 presents the results, demonstrating the superiority of Mean aggregation, while both aggregation methods surpass the performance of relying on a single image.

9.3.2 Qualitative Results

Figure 9.2 presents qualitative examples of our model's predictions, demonstrating the interpretability and plausibility of the classification-by-description approach. In the true positive prediction, the model accurately detects most descriptors and identifies "Mass in the mediastinum" as the primary indication for enlarged cardiomeastinum. For the true negative case, the model correctly detects none of the descriptors, confirming the absence of abnormal findings. However, in the false positive example, the model's mistake is easily identifiable. It detects air bronchograms with relatively high certainty but fails to detect consolidation. This error is readily apparent to a radiologist, as air bronchograms are findings that typically co-occur with consolidation (i.e., air-filled bronchi in consolidated areas). By providing the combination of descriptors that led to the decision, our approach substantially improves explainability and enables radiologists to quickly validate the model's reasoning.

In the false negative case, the model misses the presence of a pacemaker but detects some kind of implant, indicating that it understands the presence of a foreign object but cannot identify it specifically. This scenario highlights how the model's predictions, even when incorrect, can provide valuable insights into its decision-making process, facilitating error detection and interpretation by radiologists. However, this example also reveals a limitation of the current reasoning process: the probabilities of the detailed radiological findings should not simply be averaged. Instead, they should be analyzed for dependency and relevance. For instance, the presence of any foreign object should trigger the "Support Devices" class, whereas, in our method, the absence of breast implants incorrectly negates the presence of a correctly detected artificial joint. This suggests that future work should investigate more sophisticated aggregation of descriptor probabilities.

Overall, the classification-by-description approach employed by Xplainer facilitates a plausibility check of specific inference results and enhances the understanding of error sources. This aligns with the importance of explainability and interpretability in clinical decision support systems, as discussed in Section 2.2.3, enabling radiologists to validate and trust the model's predictions.

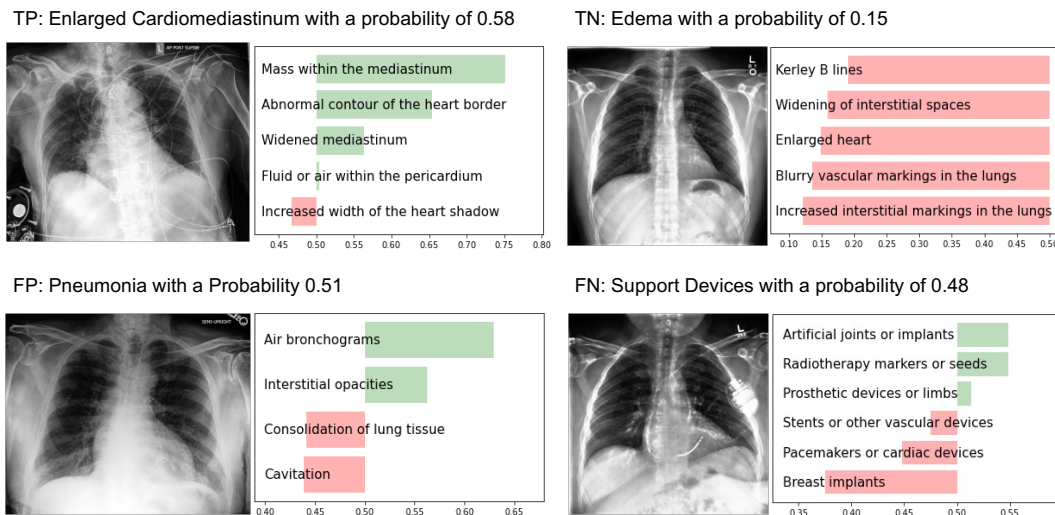


Fig. 9.2. Qualitative results of Xplainer, demonstrating the interpretability and plausibility of the classification-by-description approach. The model’s predictions, even when incorrect, provide valuable insights into its decision-making process, facilitating error detection and interpretation by radiologists. [184] *Reproduced with permission from Springer Nature.*

9.4 Discussion

One limitation of modeling a joint probability is the assumption that all descriptors appear simultaneously and have equal importance. While this simplification leads to good results, it may not always hold true, as a pathology can present with varying signs. Moreover, there may be inter-dependencies between the descriptors, where certain combinations of descriptors have a stronger correlation with the presence of a disease than others. For example, in the case of support devices, the presence of a single support device is sufficient for the label to be positive. A straightforward solution to address this is to consider only the top-k descriptor probabilities. We demonstrate experimentally on the CheXpert validation set that using only the top-1 probability can increase the AUC for support devices from 80.25% to 80.41%. Relaxing this assumption and considering the varying importance and complex relationships between descriptors is an important direction for future research.

As an initial step towards modeling descriptor importance, we explore a supervised, out-of-domain approach using a Naive Bayes classifier trained on the MIMIC-CXR dataset [118]. This classifier predicts findings using the descriptor probabilities as input, learning more complex interactions between the decomposed probabilities than simple averaging in the log space. While this requires annotation for the findings, these can be extracted from the reports using the CheXpert labeler [108]. We observe a slight performance increase on the CheXpert test set from 80.58% to 81.37% AUC, indicating that the descriptor importance learned on MIMIC can be partially transferred to an out-of-domain dataset. This finding suggests that incorporating domain knowledge and leveraging large-scale datasets can improve the generalizability and adaptability of diagnostic models, aligning with the objectives of integrating experiential and formal clinical knowledge into decision support systems discussed in Section 2.1.4. In future research, the reasoning of how each radiological finding in the image contributes to predicted

classes could also be performed by an LLM and integrated with other non-image information about the patient for holistic reasoning.

Using descriptors in Xplainer provides a flexible and adaptive approach to automated diagnosis prediction. The detection of observations using descriptions allows our model to predict a disease without relying on labeled data. This could be leveraged to adapt the model to novel or rare diseases that might be unknown or uncommon but show known symptom characteristics. This could also allow for the adaptation of the model to special patient populations, where the importance of descriptors may vary.

9.5 Conclusion

In this work, we introduced Xplainer, a novel and effective zero-shot approach for chest X-ray diagnosis prediction that explains the model's decision. Our approach makes predictions without needing label supervision by leveraging BioVil, a pretrained, domain-specific CLIP model, and employing contrastive observation-based prompting. Xplainer significantly outperformed previous zero-shot methods on the CheXpert and Chest-Xray14 datasets, demonstrating its effectiveness and potential for clinical application. Our findings emphasize the crucial role of designing informative prompts in improving model performance. The ablation studies revealed that incorporating disease indication and report style formulation into observation-based prompts substantially enhances performance, highlighting the importance of aligning prompts with the domain-specific language used in medical reports. Moreover, using contrastive prompts, which explicitly contrast positive and negative examples, significantly boosts performance. This approach mirrors the analytical reasoning process employed by clinicians, as discussed in Section 2.1.3.

We anticipate that the principles employed in Xplainer can be extended to other medical imaging domains and have practical applications in real-world clinical settings. By providing accurate and explainable diagnostic predictions, Xplainer has the potential to support clinicians in their decision-making process and improve patient outcomes. Furthermore, the flexibility and adaptability of our descriptor-based approach enable its deployment in diverse clinical contexts, including novel diseases and specific patient populations.

In conclusion, Xplainer leverages the pretraining of a vision-language model on large amounts of unstructured radiology data to enable the reasoning over structured image findings. Guided by in-domain expertise, this results in a competitive zero-shot method that provides intrinsic interpretability of its reasoning. This gives a glimpse into the future intersection between foundation models for medical domain understanding and large language models. While medical foundation models could close the gap between language and highly specialized domains like medical images or genetics, a large language model may put findings into perspective by integrating them with prior knowledge and providing interactive reasoning.

Part IV

Post-hoc Interpretation of Neural Networks

Interpretability

The final step of the deductive reasoning approach of Xplainer we discussed in Chapter 9 can be considered intrinsically interpretable due to the transparent contribution of each concept to the final prediction. However, we still need to trust the model's formation of these concepts. Moreover, it could be argued that an end-to-end optimized approach might outperform this zero-shot method given sufficient labels. Therefore, it is worth exploring post-hoc interpretation approaches that explain deep learning models that are not restricted by requirements of intrinsic interpretability - assuming there is a trade-off to be made. Returning to our analogy of an experienced clinician who can make an intuitive diagnosis instantly, it is clear that, as patients, we would still like to follow the reasoning to some extent. Moreover, the reasoning behind the decision might also be required for legal and documentation purposes. We argue that interpreting the decision-making of a well-performing deep learning model could reveal additional information about the input image beyond primary output labels [205], guiding the user's attention or pre-filling detailed structured reporting templates, as we discussed in Chapter 7.

In this part, we will provide a high-level overview of interpretability methods and then, in Chapter 11, delve into the post-hoc interpretation of a deep neural network for detecting vertebral fractures using a semantic concept activation method. This approach differs from the Xplainer we discussed in Chapter 9, where we first predicted concepts and then used them for final classification. In contrast, this method aims to discover concepts that contributed to the classification retrospectively.

Zhang et al. [270] categorize the research on the interpretability of deep learning models along three dimensions. The first distinction lies in the previously discussed difference between post-hoc methods and methods intrinsically interpretable by design. Second, various types of explanations with increasing levels of explanatory power exist. Example-based explanations, such as the prototypical networks we discussed in Section 4.1, explain their decisions by identifying relevant prototypes or examples that can be used to illustrate similar image features. Feature attributions, the most prevalent approach in medical image analysis, attempt to visualize which regions in the input image contribute the most to the classification output without providing insights into the inner workings of neural networks [123]. The next level of explanatory power focuses on understanding hidden semantics within the network. Building upon Network Dissection [15], our work described in Chapter 11 falls into this category, aiming to match activated concepts within the neural network that correspond to clinical concepts a radiologist would search for in an image. According to Zhang et al. [270], the highest level of explanatory power is offered by logic rules, such as decision trees and rules, which we have already explored in the introductory sections. Also, the generation of counterfactuals has a strong explanatory power as we have shown in our recent work on ordinal counterfactuals for the detection of vertebral body fractures using diffusion autoencoders [120]. The final distinction they provide is between explanations made on a

local input space level (an individual sample) and those made on a global level, i.e., explaining the network as a whole, such as the presence of rules or concepts. In the following chapter, we will explore both local and global explanations.

Explaining Vertebrae Fracture Detection with Semantic Concept Activations

Contents

11.1 Introduction	107
11.2 Related Work	108
11.2.1 Interpretability	109
11.3 Methodology	110
11.3.1 Vertebral Fracture Detection	110
11.3.2 Extraction of Semantic Concepts	110
11.3.3 Concept Correlation at Inference	111
11.4 Experimental Setup	111
11.5 Results and Discussion	112
11.5.1 Vertebral Fracture Detection	112
11.5.2 Clinical Evaluation of Semantic Concepts	113
11.5.3 Single-Inference Concept Visualization	114
11.6 Conclusion	115

11.1 Introduction

Osteoporosis is a prevalent disease in the elderly population, affecting millions of individuals worldwide [36, 95]. The early detection of incidental osteoporotic fractures in routine computed tomography (CT) scans is crucial, as these fractures often remain asymptomatic for an extended period [85]. Moreover, osteoporotic fractures are independent predictors of subsequent fractures, with a significantly increased risk and mortality rate [37, 158]. The consequences of osteoporotic fractures include substantial socioeconomic impacts and a diminished quality of life for affected individuals [23, 38, 88, 111]. Despite the clinical importance of these fractures, a significant proportion of osteoporotic fractures are not adequately reported in radiological findings of routine CT scans, potentially due to the increasing workload of radiologists [12, 246].

In the older population, distinguishing between osteoporotic and malignant fractures is essential, as they have different prognoses and treatment approaches. However, differentiating between these fracture types at the vertebral body level using standard imaging techniques is often challenging and may require further diagnostic procedures, such as biopsies [68, 257]. Applying deep learning models for automated fracture detection can address these issues and

lead to standardized fracture classifications, which radiologists currently use inconsistently. However, as discussed in Section 2.2.3, most deep learning models are black-box systems that do not provide insights into their decision-making processes. Exploring the internal workings of these models can enable the investigation of failure cases and, when addressed, enhance the robustness and trustworthiness of the system. Recent research has revealed that neural networks trained for classification tasks can learn abstract semantic concepts similar to the patterns used by humans to differentiate images [15]. If the concepts discovered in vertebral fracture classification align with clinicians' image features, they could also be employed to determine fracture types and generate automatic image descriptions and diagnostic reasoning for report generation. This approach is similar to an experienced clinician making an intuitive assessment (System 1, inductive reasoning - see Section 2.1.3) of a fracture type and then retrospectively looking for radiological findings that support his conclusions, either writing this in a report or explaining his diagnosis to a patient.

Interpretable diagnosis has primarily been investigated using feature attribution (saliency) approaches [123], such as class activation maps [272]. These interpretations reveal the location of important features for prediction. While feature attribution is a valuable tool for verifying the network's inference mechanism, it does not provide additional information regarding the prediction. Furthermore, knowing only the location of important features is not helpful for fracture diagnosis, as it is easy to identify the fracture location. Instead, it is more interesting to understand "what" features are important for the diagnosis.

Drawing inspiration from the network dissection technique [15] and its applications in chest radiography [123] and mammography [252], in our work [71], we analyze the inner workings of a neural network and the correspondence to semantic concepts on both a local input space and global level. In the global setting, we compute the output of the last convolutional layer for all input data and identify neurons that exhibit a strong correlation with the output value associated with fractures. We then ask clinicians to identify the concepts associated with highly correlated activations by examining the inputs most strongly activating those neurons. This setting provides a comprehensive understanding of the concepts learned by the network and whether they align with the concepts used by clinicians. In the single-inference setting, we identify the highly activated convolutional neurons for a single input and visualize their associated concepts by presenting the top images that activate each neuron. This setting enables users to gain a conceptual understanding of the model's decision-making process. We analyze both settings using the open-source VerSe [211] dataset and a larger private dataset from our hospital. These concept-based interpretations serve as a foundation for the broader goal of explainable diagnosis and the generation of radiology reports. The primary objectives of this work are to investigate the features utilized by the network for fracture diagnosis, determine their overlap with clinical knowledge, and explore how they can be employed to enhance the verbosity and explainability of fracture diagnosis.

11.2 Related Work



Contributions:

- We propose a method to identify activated neurons in 3D CNN that are highly correlated with vertebral fracture detection, enabling the assessment of their correspondence to clinical concepts.
- We qualitatively evaluate the identified concepts with medical experts.
- We introduce a visual explanation approach for the network's decision-making process by highlighting the most relevant concepts for classifying a given sample.

Vertebral Fracture Detection

Numerous approaches have been proposed for automatically detecting vertebral fractures. Most of these methods employ Convolutional Neural Networks (CNNs) on Computer Tomography (CT) spine images. However, there are notable exceptions, such as [232], which utilizes tabular data extracted from CT images in a Random Forest, and [47, 167], which focuses on detecting fractures in X-rays.

Both 2D and 3D methods have been explored for fracture detection. 2D methods often rely on feature aggregation using Recurrent Neural Networks to model inter-slice dependencies [11, 231]. Hussein et al. [107] operate on reformatted 2D slices in the sagittal view and perform fracture grading using a specialized loss. Pisov et al. [191] perform a key point detection for measuring the compression of each vertebra, using this for both fracture grading and detection. Nicolaes et al. [169] pioneered using 3D convolutions for vertebral fracture detection, focusing on detecting fractures at the voxel level and then post-processing the results. Chettrit et al. [50] proposed to model the inter-volume dependencies with a sequential model, and [261] employed a 3D model to detect osteoporotic fractures on a patient level.

In addition to fracture detection and grading, recent studies by Li et al. [143] and Feng et al. [73] have explored the distinction between benign and malignant vertebral fractures. Despite the growing body of research on vertebral fracture detection, the interpretability of these models remains unexplored mainly, except Yilmaz et al. [262], which investigated the usefulness of attribution maps in osteoporotic fracture discrimination. Our work differs methodologically from [15, 123] because we do not utilize an annotation dataset. Instead, we identify neurons that correlate highly with the output under investigation. We investigate a different medical domain and explore distinct research questions, such as the features contributing to true and false positives.

11.2.1 Interpretability

The interpretability of models in the domain of vertebral fracture diagnosis has been explored to a limited extent. Yilmaz et al. [262] interpret models using feature attribution (saliency) approaches to identify regions in the input contributing to the prediction. Feature attribution is the predominant approach in most medical image analysis applications [123]. However, these

methods have limitations in revealing information about the model’s decision-making mechanism. Furthermore, the feature attribution problem remains largely unsolved, and despite the existence of various attribution approaches (e.g., CAM [272], LRP[164], DeepSHAP[152], IBA [208, 268]), there is often disagreement among the methods regarding the identified important features [268]. This disagreement poses a challenge for domain experts utilizing these attribution methods. Consequently, there is a need for reliable interpretation approaches that provide more information beyond simply identifying "which region is important." The Network Dissection approach [15] offers a different approach by identifying the concepts encoded by the network’s internal units (neurons). Drawing motivation from this approach, Wu et al. [252] identify the concepts encoded by networks for diagnosis in mammography images. In contrast, Khakzar et al. [123] perform dissection on chest x-ray models and investigate research questions such as the clinical concepts networks capture when trained on COVID-19 severity scores.

11.3 Methodology

11.3.1 Vertebral Fracture Detection

The vertebral fracture detection task is formulated as a binary classification problem, where the positive class indicates the presence of a fracture. The network function is represented as $f_{\Theta}(x) : \mathbb{R}^{H \times W \times D} \rightarrow \mathbb{R}$, and the predicted probability is obtained by applying the sigmoid function to the network output, i.e., $\hat{y} = \text{sigmoid}(f_{\Theta}(x))$. We employ a 3D U-Net architecture [51] for the vertebral fracture classification task, modifying its upsampling path by replacing it with a fully connected layer for classification similar to the image feature extraction described in Section 6.3 and [121] sans the decoder used for segmentation. Preliminary experiments with a DenseNet [102] yielded similar classification performance (0.937 AUC compared to 0.933 for the 3D-UNet) for this task. However, due to its erratic learning curves and inferior results in downstream tasks, such as detecting vertebral fractures in public datasets, we opted against using this architecture. Binary Cross Entropy (BCE) is the loss function for training the model.

11.3.2 Extraction of Semantic Concepts

In neural networks, each neuron responds to a particular input pattern, and this response pattern can be interpreted as the neuron’s associated *concept*. In the context of convolutional neural networks (CNNs), a neuron can be considered an entire activation map or an individual activation unit within the map. Given that all activation units within a single activation map serve the same function, differing only in their spatial locations, they collectively represent a single concept [15]. This property allows us to treat an activation map as a unified representation of a specific concept, enabling the analysis and interpretation of the learned features within the CNN.

Let $A \in \mathbb{R}^{H' \times W' \times K}$ represent the tensor of output activations from the network’s final convolutional layer, where H' and W' denote the spatial dimensions, and K signifies the

number of channels in this layer. To identify the most relevant units, we first analyze the distribution of activations for each unit a_k . We then establish a threshold \mathcal{T}_k for each unit k such that the probability of its activation exceeding this threshold is 0.005, i.e., $P(a_k > \mathcal{T}_k) = 0.005$ [15]. Based on these thresholds, we construct a binary segmentation mask $M_k(\mathbf{x})$ for each unit k and input \mathbf{x} , defined as $M_k(\mathbf{x}) := A_k(\mathbf{x}) > \mathcal{T}_k$. This mask indicates whether the activation of unit k exceeds its corresponding threshold for the given input. Finally, we define the set of enabled units E_x for an input \mathbf{x} as $E_x := \{k \mid \sum M_k(\mathbf{x}) > 0\}$, which includes all units whose activations exceed their respective thresholds for the input \mathbf{x} .

Positive Prediction Correlation

Certain units within the network may encode concepts that are particularly informative for determining whether a sample is fractured. These units correlate more strongly with true positive predictions than others. To identify such units, we introduce positive prediction correlation as:

$$c_k := \frac{\sum_{x \in P} \mathbb{1}_{E_x}(k)}{|P|} \quad (11.1)$$

where P represents the set of all positive samples, and $\mathbb{1}$ denotes the indicator function. By ranking the units based on their positive prediction correlation values, we can identify the units that are most strongly associated with true positive predictions. For instance, if $c_{k_1} > c_{k_2} > \dots$, then unit k_1 exhibits the highest correlation with true positive predictions, followed by unit k_2 , and so on.

11.3.3 Concept Correlation at Inference

Given the diverse nature of defects observed in fractured vertebrae, varying concepts may be relevant during inference for different samples. To compute the relevance of each unit k in the context of a specific input \mathbf{x} , we introduce a measure called inference relevance as follows:

$$r_k := \sum M_k(\mathbf{x}) \odot A_k(\mathbf{x}) \quad (11.2)$$

For example, if $r_{k_1} > r_{k_2}$ for units k_1 and k_2 , then unit k_1 is more relevant than unit k_2 for the inference of \mathbf{x} . To visualize the highly correlating concepts for a specific sample \mathbf{x} , we first compute the inference relevance for each unit and then display the activation maps $A_{k_1}(x)$, $A_{k_2}(x)$, and so on, in descending order of their inference relevance values.

11.4 Experimental Setup

Data Preparation

We train the network on two datasets: the VerSe dataset [211] and an in-house dataset collected at Klinikum rechts der Isar and Klinikum der Universität München in Munich,

Germany. The in-house dataset includes 465 patients with a median age of 69 (± 12) years. It contains a diverse collection of CT scans with varying fields of view, scanner settings, and a mix of healthy and fractured vertebrae, including cases with metallic implants and foreign materials. Combining both datasets creates a comprehensive collection of CT images featuring healthy and fractured vertebrae with osteoporotic or malignant fractures acquired using different CT scanners. To mitigate the class imbalance inherent in the data, we employ undersampling of negative samples and oversampling of positive (fractured) samples during training. This approach ensures a balanced class distribution within each training epoch. As osteoporotic and malignant fractures are relatively rare in cervical vertebrae (C1-C7), these vertebrae are excluded from the dataset.

For each vertebra, we extract 3D patches of size $96 \times 96 \times 96$ with a resolution of 1mm. These patches are centered on the vertebral body and aligned along the spine by orienting the vertical axis with a spline constructed using the vertebral centroids provided by the dataset, following an approach similar to [107]. The intensity values of the resulting patches are cropped to a Hounsfield Unit range of $[-1000, 1000]$ and then scaled to $[0, 1]$. We apply data augmentation during training to enhance the network's robustness and generalization ability. These augmentations include intensity transformations (Gaussian noise, smoothing, and contrast adjustment) and substantial spatial transformations (similarity transformation and elastic deformation). The data preprocessing and augmentation steps are implemented using the NiBabel 3.2.1 and MONAI 0.8.0 libraries.

Implementation Details

The 3D U-Net is implemented using PyTorch Lightning 1.5.10 and PyTorch 1.10.2. The model is trained with the Adam optimizer, employing a learning rate of 0.001 without weight decay. The training process is terminated if the validation F1 score does not improve for 50 epochs, serving as an early stopping criterion. Dropout regularization with a probability of 0.3 is applied during training.

11.5 Results and Discussion

Before we dissect the activated semantic concepts, we assess our model's predictive performance in detecting vertebral fractures. Then, we qualitatively evaluate the clinical soundness of extracted concepts and how these concepts affect an individual inference.

11.5.1 Vertebral Fracture Detection

To assess the performance of our vertebral fracture detection model, we employ threshold-dependent evaluation metrics, namely F1-score and accuracy. Additionally, we consider threshold-independent metrics such as the area under the curve (AUC) and average precision (AP). We conduct five separate training runs for each model and report the mean and standard deviation of the metrics.

Tab. 11.1. Evaluation of the trained neural networks' performance on the test holdout of the VerSe dataset and the combined dataset, which includes VerSe and proprietary data obtained from Klinikum rechts der Isar and Klinikum der Universität München. The VerSe dataset consists of 3,920 non-cervical vertebrae, with 254 fractures, while the combined dataset encompasses 10,675 T1-L5 vertebrae, including 1,246 fractures. [71] *Reproduced with permission from Springer Nature.*

Training	Testing	F1 (%)	Acc. (%)	AUC (%)	AP (%)
VerSe	VerSe	71.2 ± 10.8	78.2 ± 12.0	84.5 ± 9.1	76.4 ± 14.5
VerSe, in-house	VerSe	86.1 ± 2.6	90.9 ± 1.6	96.2 ± 0.9	94.1 ± 1.6
VerSe, in-house	VerSe, in-house	88.0 ± 0.7	88.0 ± 0.4	94.7 ± 0.5	95.0 ± 0.4

Networks trained solely on the VerSe dataset demonstrate performance comparable to simplistic 2D vertebral fracture detection methods applied to the same dataset [107]. However, these networks highly rely on favorable random seed initialization and fail to produce detector units with distinctive patterns. To overcome these limitations, we train a network using an expanded dataset that combines VerSe with proprietary data acquired from Klinikum rechts der Isar and Klinikum der Universität München. As shown in Table 11.1, this network consistently performs better and yields detector units exhibiting diverse patterns. The following sections will thoroughly examine the characteristics and implications of these patterns.

11.5.2 Clinical Evaluation of Semantic Concepts

After training the network on the expanded dataset, we employ an extended version of Network Dissection [15] to extract its semantic concepts in 3D space. To focus on the most informative units, we select the top ten detector units that highly correlate with true positive predictions, as described in Section 11.3.2. We generate a single-slice collage of 25 highly activating fractured samples for each unit, providing an overview of the units' activation patterns. Additionally, we export all 2D slices and three-dimensional NIFTI files for the five samples with the highest activation levels, enabling a comprehensive analysis.

To assess the clinical relevance of these detector units, we consult two clinical experts with a combined experience of 22 years in spine imaging. Excluding three units lacking immediate associations, we present the remaining detector units in Figure 11.1, ranked by their correlation and corresponding clinical explanations. The provided samples showcase a diverse range of detector unit activations, with each unit demonstrating consistent patterns across multiple samples. These units primarily focus on the main vertebra, even if some activation occurs in the surrounding regions. The observed patterns align with the bone anatomy and manifest in clinically significant locations. Since severe fractures are associated with superior and inferior vertebral endplate changes, most activations are found in these areas. Although multiple detector units target these regions, they concentrate on different locations and exhibit varying sizes of regions of interest, with some units incorporating additional information from the intervertebral discs and adjacent vertebrae. These findings are clinically meaningful for detecting moderate and severe vertebral deformations (Genant grade 1 or higher [78]), indicating that our network has learned concepts with clinical relevance.

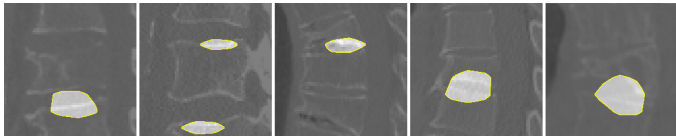
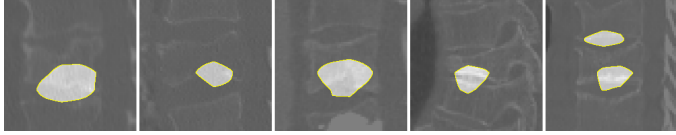
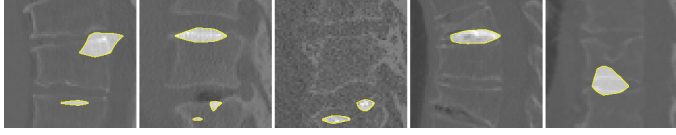
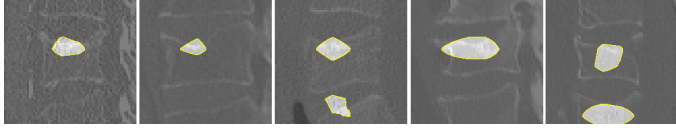
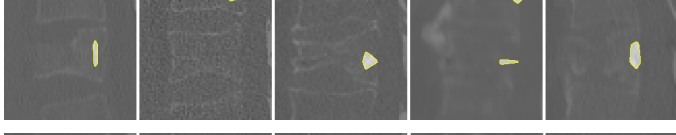
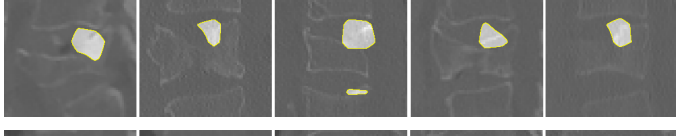
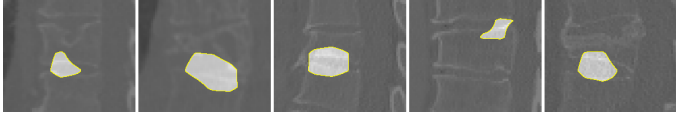
Rank	Sample Activations	Clinical Concepts
1		Abnormal endplate and intervertebral disc shapes
2		Primarily defects of the inferior endplate, associated with severe fractures
5		Abnormal endplate shapes with partial observation of adjacent inferior vertebrae
7		Central defect of the superior endplate, commonly observed in compression fractures, with partial observation of adjacent inferior vertebrae
8		Observation of the spongiosa in the primary vertebrae as well as the adjacent superior one
9		Injury to the middle column of the vertebral bodies, associated with clinically significant myelon compression and consecutive paresis
10		Abnormal endplate and intervertebral disc shapes

Fig. 11.1. Visualization of the detector units most highly correlated with a true positive prediction and clinical experts' interpretation of their activations. All displayed samples are fractured and represented by a slice with high activation after thresholding. [71] *Reproduced with permission from Springer Nature.*

For the omitted cases, we observe either statistically insignificant activations (i.e., $M_k(\mathbf{x}) = \mathbf{0}$) or sporadic activations that lack clear patterns, despite their high correlation with true positive predictions. However, such detector units constitute a minority and can be disregarded in favor of those exhibiting tangible patterns.

11.5.3 Single-Inference Concept Visualization

Having validated the network's ability to learn clinically relevant concepts, we aim to provide further insight into its decision-making process by offering visual explanations for individual inferences. To achieve this, we propose a system that visualizes the concepts deemed most important by the network during inference.

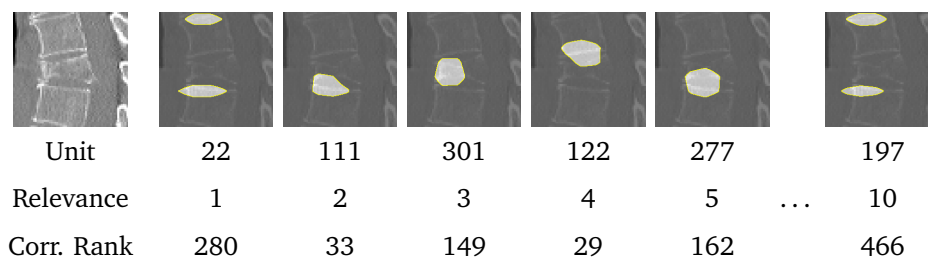


Fig. 11.2. Visualization of the most relevant detector units during the classification of the sample shown on the left, which the network correctly identifies as fractured. Each detector unit is represented by a single slice activation for that specific sample, and its ranking is based on its high correlation with true positive predictions. The visualization reveals that the network uses concepts associated with wedge-shaped deformity and incorporates information from an adjacent vertebra. [71] *Reproduced with permission from Springer Nature.*

Utilizing the method described in Section 11.3.3, we identify the units representing the most relevant concepts and retrieve their corresponding top activating images from our combined dataset. For each unit, we present two visualizations: (i) activations for the input sample and (ii) the activations for its top images. This approach shows the user a detector unit’s response for a given input image and the concept activation in a larger context. Both visualizations display a single slice with high activation levels after thresholding. Figure 11.2 gives an example for (i), demonstrating the network’s use of diverse concepts like the wedge-shaped deformity and the incorporation of information from an adjacent vertebra.

This visualization allows users to understand the network’s decision-making process better and identify relevant semantic concepts. This can build trust in the system, detect biases, and potentially be used to extract more granular findings than the supervision signal of fractures.

11.6 Conclusion

In this work, we showed that a 3D CNN can learn a diverse set of concepts to detect vertebral body fractures. To evaluate the relevance of these concepts, we introduce a method to identify concepts that highly correlate with fracture detection. We then evaluated how these discovered concepts overlap with clinically relevant semantic concepts and described the discovered concepts with clinical experts. Finally, we proposed a visualization method that displays the concepts used for an individual sample and explored how this can give the user insight into the decision-making process of the model.

In future research, semantically similar concepts could be grouped with annotations or text supervision and leveraged for the extraction of granular radiological findings in structured reporting or for providing semantically meaningful explanations in the form of concept activations. In the context of clinical reasoning, this approach demonstrates the post-hoc interpretation of a data-driven system similar to the retrospective explanation of an intuitive decision that may have come into question. Such systems might be used to gain a deeper image

understanding when confronted with uncertain predictions that require further assessment of detailed image features and their integration with other modalities.

Part V

Conclusion and Outlook

Conclusion

This thesis set out to explore the potential of advancing multimodal deep learning methods to support clinicians in their clinical decisions and reasoning. After detailing the clinical decision process (Chapter 2) and laying out the fundamentals of multimodal deep learning (Chapter 3), we addressed this challenge threefold: integration of multimodal patient data to create comprehensive patient representations, incorporation of prior knowledge into deep learning models to leverage existing medical expertise, and improving the understanding of the decision-making process of deep learning models, enhancing their interpretability and trustworthiness.

Modelling Multimodal Patient Data and Attention-based Reasoning

U-GAT (Chapter 6) combines approaches from a range of deep learning research topics, including multimodal learning, multitask learning, and graph deep learning on population graphs, to solve a challenging problem of COVID-19 outcome prediction with limited data. In addition, we segmented lung pathologies in the input CT images, which served as an auxiliary supervision signal and allowed us to extract tabular image information in the form of radiomics. This enabled us to integrate the modalities on two different levels of abstraction: tabular data (radiomics) and latent representations. Combining radiomics with clinical patient information allows us to model the inter-patient relationship in a population graph that was further processed using multimodal latent features for each node. By clustering previous patients in a graph based on their outcomes and selecting the most relevant patients using an attention mechanism, we model clinicians' retrieval process when making intuitive decisions based on their recollection of similar patients. By visualizing the attention and graph neighborhood, we can provide insight into this inductive reasoning of the deep learning model.

Incorporation of Prior Formal Knowledge in the Reasoning Process

In ToxNet (Chapter 5), we demonstrated that integrating both the database of previous patients and matching patient symptoms with clinical literature improves the prediction of intoxication. This approach showcases the value of combining experiential with formal knowledge in decision support systems. The improved performance highlights that such decision support systems can be effective in high-stakes, time-critical situations, as in an emergency hotline setting where clinicians may not have the time to consider all available data or resort to analytical reasoning based on literature. In Xplainer (Chapter 9), we showed that the classification-by-description approach provides intrinsic interpretability and is a highly effective zero-shot method. This method outperformed previous approaches that were already competitive with radiologists, demonstrating the power of integrating prior formal knowledge into the reasoning process.

Reasoning over Fine-grained Semantic Image Concepts

We viewed reasoning over semantic concepts in radiological images from two perspectives: the top-down, analytical approach of Xplainer, where we test the image for the presence of predefined image observations, and the bottom-up approach of concept activations (Chapter 11). Here, we directly predict high-level image findings, such as vertebral fractures, and retrospectively analyze the neural activations for semantic concepts. The extraction of detailed and structured concepts from a diagnostic image is highly relevant for several reasons: First, these fine-grained findings are crucial for the holistic decision-making process since they can be put into perspective with other non-imaging biomarkers to support or reject a hypothesis formed solely on image information. Second, such semantic concepts could match the elements of standardized reports, which clinicians use to document their reasoning for radiological findings in structured reporting. In the future, this reasoning could facilitate an interactive reporting process. Third, reasoning correctly over low-level image concepts to arrive at a diagnostic decision is a powerful metric for assessing a clinical decision support system's image understanding, which clinicians can intuitively verify.

In summary, this thesis contributes to advancing holistic clinical decision support by addressing the integration of multimodal patient data, the incorporation of prior knowledge, and the improvement of model interpretability. By exploring novel approaches that align with clinicians' cognitive processes and provide insights into the reasoning behind the models' decisions, we have taken essential steps toward creating trustworthy and effective clinical reasoning support systems.

Outlook

The field of deep learning is advancing at an unprecedented rate. During the final stages of this research, a new paradigm emerged: large-scale foundation models, particularly highlighted by generative models for images and text. In language tasks such as clinical text summarization [233], large language models (LLMs) have already surpassed human performance and demonstrated their ability to pass medical exams [173]. Furthermore, our recent works on multimodal LLMs on conversational radiology reporting [186] and knowledge-guided surgery understanding [176] have demonstrated that these foundation models can be adapted to highly specialized domains. However, given the complexity and multimodal nature of clinical decision-making, the application of LLMs in clinical reasoning remains a challenge [86]. This section provides an outlook on how the findings from this thesis could impact multimodal LLMs and outlines potential pathways for future research.

Clinical Reasoning on Multimodal Data in Large Language Models

The beauty of operating in the language space is that we can naturally describe the quintessence of most other modalities relevant to clinical decision-making since this is how medical doctors communicate. A large amount of patient information, such as the chief concern, patient history, and various reports, is already expressed as text within the electronic health record (EHR). Tabular data can be described by categories or binned and relatively easily expressed in words. For example, a blood value could be described as normal, low, extremely high, etc., depending on its relevance to the clinical decision at hand. As discussed in Sections 2.1.2 and 3.1, this tabular information can also include a variety of diagnostic tests and genetic information. For medical images, as demonstrated in RaDialog [186] and described earlier, fine-grained findings can be extracted as standardized or unstructured reports, thereby expressing visual information in language.

However, as emphasized in Section 2.1.2, clinical observations are not black and white, and every finding is associated with uncertainty. Since integrating text in the context prompt of LLMs is essentially a form of early fusion (see Section 3.3.1), all the disadvantages of this method apply, such as the lack of interaction during feature extraction and the absence of task-specific optimization. To address these challenges, we must find ways to express uncertainty on different abstraction levels and understand how it propagates through the decision process. One remedy to this is encoding images, as done in RaDialog, or tabular data, as done in HeLM [18], in the token space of a large language model and performing an end-to-end optimization resembling a joint fusion strategy as explored in this work.

For effective use of multimodal data, an LLM could be trained to operate as an agent with the objective of reducing the uncertainty of the diagnostic differential by asking for missing diagnostic tests and continuously integrating new information until a confident diagnosis or treatment recommendation can be made. This approach would enable the LLM to actively sup-

port the clinical reasoning process, mimicking how clinicians gather and analyze information to arrive at a diagnosis.

Foundation Models and the Integration of Experiential and Formal Knowledge

The breakthroughs of foundation models have largely been achieved by using large-scale, publicly available non-medical data. However, if we overcome the challenges of interoperability and data protection [193] and manage to train foundation models on large-scale medical data, these models could become a powerful way to create multimodal patient representations. The contrastive models explored in this thesis are just a precursor to future developments when larger amounts of data and many different patient modalities could be used for pretraining, as opposed to just radiology reports and images. ImageBind [79] has demonstrated that multiple modalities can be joined in this process, and recent generative models like CoDi [227] and Emu [220] have shown the potential of integrating multiple modalities both as input and generated output. Such holistic patient representations could be used for the integration of exemplar knowledge by retrieving relevant embeddings and then providing text information for further processing in an LLM, similar to the retrieval in CXR-RePaiR [69].

Formal knowledge can be incorporated in a straightforward manner by retrieving an in-depth article about a rare disease or by retrieving a clinical guideline for a common disease. This can be implemented using retrieval augmented generation (RAG) [140]. An alternative to RAG and unstructured knowledge is the use of knowledge graphs and reasoning on graphs (ROG), as introduced by Luo et al. [153]. These approaches provide powerful tools for sampling relevant and discrete knowledge about a particular problem. However, ensuring that the reasoning follows these guidelines remains a challenge. While rule-based systems offer high interpretability, they often fall short in performance, as we have discussed earlier. This dilemma naturally leads us to the next question: how can we model both intuitive and analytical reasoning to balance interpretability and performance?

Mimicking the Dual System for Clinical Reasoning Support

A key insight of this thesis is that both intuitive (data-driven) and analytical (knowledge-driven) types of reasoning have their place in a clinician's mind and a machine's algorithm. For future systems, we should not only consider the ends of the continuum but also their interplay. An experienced doctor will make most decisions intuitively, particularly under time pressure, and an automated system should do the same using data-driven principles. However, when faced with high uncertainty, like with rare diseases or contradicting diagnostic tests, these systems could interactively support medical doctors in reasoning by compensating for the limitations of the human mind, such as processing vast amounts of experiential data and finding patterns in a multitude of complex patient data.

As these collaborative methods continue to evolve and be refined, they have the potential to significantly improve patient outcomes and bring us closer to making healthcare accessible and personalized for everyone.

Part VI

Appendix

Authored and Co-authored Publications

First Author

1. **M. Keicher**, K. Zaripova, T. Czempiel, K. Mach, A. Khakzar, and N. Navab. “FlexR: Few-shot Classification with Language Embeddings for Structured Reporting of Chest X-rays”. In: *International Conference on Medical Imaging with Deep Learning – MIDL 2023*. (Full paper). Proceedings of Machine Learning Research, July 2023.
2. **M. Keicher***, M. Atad*, D. Schinz, A. S. Gersing, S. C. Foreman, S. S. Goller, J. Weissinger, J. Rischewski, A.-S. Dietrich, B. Wiestler, J. S. Kirschke, and N. Navab. “Semantic Latent Space Regression of Diffusion Autoencoders for Vertebral Fracture Grading”. In: *arXiv Preprint 2303.12031, presented at iMIMIC 2023 - Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2023*. (*Equal contribution. **Best paper award.**) Mar. 2023.
3. **M. Keicher***, H. Burwinkel*, D. Bani-Harouni*, M. Paschali, T. Czempiel, E. Burian, M. R. Makowski, R. Braren, N. Navab, and T. Wendler. “Multimodal graph attention network for COVID-19 outcome prediction”. In: *Scientific Reports* 13.1 (Nov. 2023). (*Equal contribution.)
4. C. Pellegrini*, **M. Keicher***, E. Özsoy, and N. Navab. “Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Lecture Notes in Computer Science. (*Equal contribution.) Springer Nature Switzerland, Oct. 2023, pp. 409–419.
5. C. Pellegrini*, **M. Keicher***, E. Özsoy*, P. Jiraskova, R. Braren, and N. Navab. “Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Lecture Notes in Computer Science. (*Equal contribution.) Springer Nature Switzerland, Oct. 2023, pp. 420–429.
6. P. Engstler*, **M. Keicher***, D. Schinz, K. Mach, A. S. Gersing, S. C. Foreman, S. S. Goller, J. Weissinger, J. Rischewski, A.-S. Dietrich, B. Wiestler, J. S. Kirschke, A. Khakzar, and N. Navab. “Interpretable Vertebral Fracture Diagnosis”. In: *Interpretability of Machine Intelligence in Medical Image Computing*. Lecture Notes in Computer Science. (*Equal contribution.) Springer Nature Switzerland, 2022, pp. 71–81.
7. M. Kollovich*, **M. Keicher***, S. Wunderlich, H. Burwinkel, T. Wendler, and N. Navab. “U-PET: MRI-based Dementia Detection with Joint Generation of Synthetic FDG-PET Images”. In: *arXiv Preprint 2206.08078*. (*Equal contribution.) June 2022.
8. A. Sankar*, **M. Keicher***, R. Eisawy, A. Parida, F. Pfister, S. T. Kim, and N. Navab. “GLOWin: A Flow-based Invertible Generative Framework for Learning Disentangled Feature Representations in Medical Images”. In: *arXiv Preprint 2103.10868*. (*Equal contribution.) Mar. 2021.

9. H. Burwinkel*, **M. Keicher***, D. Bani-Harouni*, T. Zellner, F. Eyer, N. Navab, and S.-A. Ahmadi. “Decision Support for Intoxication Prediction Using Graph Convolutional Networks”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Lecture Notes in Computer Science. (*Equal contribution.) Springer International Publishing, Oct. 2020, pp. 633–642.

Last Author

1. L. Buess, M. F. Stollenga, D. Schinz, B. Wiestler, J. Kirschke, A. Maier, N. Navab, and **M. Keicher**. “Video-CT MAE: Self-supervised Video-CT Domain Adaptation for Vertebral Fracture Diagnosis”. In: *Accepted for presentation at Medical Imaging with Deep Learning*. 2024.
2. Y. Xiong, J. Liu, K. Zaripova, S. Sharifzadeh, **M. Keicher***, and N. Navab*. “Prior-RadGraphFormer: A Prior-Knowledge-Enhanced Transformer for Generating Radiology Graphs from X-Rays”. In: *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology*. Ed. by S.-A. Ahmadi and S. Pereira. (*Equal contribution.) Cham: Springer Nature Switzerland, 2024, pp. 54–63.
3. C. Pellegrini*, E. Özsoy*, B. Busam, N. Navab, and **M. Keicher**. “RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance”. In: *arXiv Preprint 2311.18681*. (*Equal contribution.) Nov. 2023.

Co-Author

1. E. Özsoy*, C. Pellegrini*, **M. Keicher**, and N. Navab. “ORacle: Large Vision-Language Models for Knowledge-Guided Holistic OR Domain Modeling”. In: *arXiv Preprint 2404.07031*. (*Equal contribution.) 2024.
2. T. Zellner, K. Romanek, C. Rabe, S. Schmoll, S. Geith, E.-C. Heier, R. Stich, H. Burwinkel, **M. Keicher**, D. Bani-Harouni, N. Navab, S.-A. Ahmadi, and F. Eyer. “ToxNet: an artificial intelligence designed for decision support for toxin prediction”. In: *Clinical Toxicology* 61.1 (Jan. 2023), pp. 56–63.
3. M. Atad, V. Dmytrenko, Y. Li, X. Zhang, **M. Keicher**, J. Kirschke, B. Wiestler, A. Khakzar, and N. Navab. “CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN”. In: *arXiv Preprint 2207.07553, presented at the workshops Interpretable Machine Learning in Healthcare at ICML 2022 and Explainable AI for Computer Vision at CVPR 2022*. July 2022.
4. A. Bitarafan, M. F. Azampour, K. Bakhtari, M. Soleymani Baghshah, **M. Keicher**, and N. Navab. “Vol2Flow: Segment 3D Volumes Using a Sequence of Registration Flows”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Lecture Notes in Computer Science. Springer Nature Switzerland, 2022, pp. 609–618.
5. T. Czempiel, C. Rogers, **M. Keicher**, M. Paschali, R. Braren, E. Burian, M. Makowski, N. Navab, T. Wendler, and S. T. Kim. “Longitudinal Self-Supervision for COVID-19 Pathology Quantification”. In: *arXiv Preprint 2203.10804*. Mar. 2022.

6. D. M. Hedderich, **M. Keicher**, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, T. Czempiel, J. E. Spiro, D. Pinto dos Santos, D. Heim, C. Zimmer, D. Rückert, J. S. Kirschke, and N. Navab. “AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging”. In: *Healthcare* 9.10 (Oct. 2021), p. 1278.
7. S. T. Kim*, L. Goli*, M. Paschali, A. Khakzar, **M. Keicher**, T. Czempiel, E. Burian, R. Braren, N. Navab, and T. Wendler. “Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Lecture Notes in Computer Science. (*Equal contribution.) Springer International Publishing, 2021, pp. 273–282.
8. T. Czempiel, M. Paschali, **M. Keicher**, W. Simson, H. Feussner, S. T. Kim, and N. Navab. “TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 343–352.
9. A. Elskhawy, A. Lisowska, **M. Keicher**, J. Henry, P. Thomson, and N. Navab. “Continual Class Incremental Learning for CT Thoracic Segmentation”. In: *Workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 106–116.

Abstracts of Publications not Discussed in this Thesis

Video-CT MAE: Self-supervised Video-CT Domain Adaptation for Vertebral Fracture Diagnosis

L. Buess, M. F. Stollenga, D. Schinz, B. Wiestler, J. Kirschke, A. Maier, N. Navab, and **M. Keicher**, *Accepted for presentation at Medical Imaging with Deep Learning, 2024.*

Early and accurate diagnosis of vertebral body anomalies is crucial for effectively treating spinal disorders, but the manual interpretation of CT scans can be time-consuming and error-prone. While deep learning has shown promise in automating vertebral fracture detection, improving the interpretability of existing methods is crucial for building trust and ensuring reliable clinical application. Vision Transformers (ViTs) offer inherent interpretability through attention visualizations but are limited in their application to 3D medical images due to reliance on 2D image pretraining. To address this challenge, we propose a novel approach combining the benefits of transfer learning from video-pretrained models and domain adaptation with self-supervised pretraining on a task-specific but unlabeled dataset. Compared to naive transfer learning from Video MAE, our method shows improved downstream task performance by 8.3 in F1 and a training speedup of factor 2. This closes the gap between videos and medical images, allowing a ViT to learn relevant anatomical features while adapting to the task domain. We demonstrate that our framework enables ViTs to effectively detect vertebral fractures in a low data regime, outperforming CNN-based state-of-the-art methods while providing inherent interpretability. Our task adaptation approach and dataset not only improve the performance of our proposed method but also enhance existing self-supervised pretraining approaches, highlighting the benefits of task-specific self-supervised pretraining for domain adaptation. The code for our method and dataset creation is publicly available.

ORacle: Large Vision-Language Models for Knowledge-Guided Holistic OR Domain Modeling

E. Özsoy*, C. Pellegrini*, **M. Keicher**, and N. Navab, *arXiv Preprint 2404.07031, 2024* (*Equal contribution.)

Every day, countless surgeries are performed worldwide, each within the distinct settings of operating rooms (ORs) that vary not only in their setups but also in the personnel, tools, and equipment used. This inherent diversity poses a substantial challenge for achieving a holistic understanding of the OR, as it requires models to generalize beyond their initial training datasets. To address this challenge, we introduce ORacle, an advanced vision-language model designed for holistic OR domain modeling. ORacle incorporates multi-view and temporal capabilities and can leverage external knowledge during inference, enabling it to adapt to previously unseen surgical scenarios. This capability is further enhanced by our novel data augmentation framework, which significantly diversifies the training dataset, ensuring ORacle's proficiency in applying the provided knowledge effectively. In rigorous testing, including scene graph generation and downstream tasks on the 4D-OR dataset, ORacle not only demonstrates state-of-the-art performance but does so requiring less data than existing models. Furthermore, its adaptability is displayed through its ability to interpret unseen views, actions, and appearances of tools and equipment. This demonstrates ORacle's potential to significantly enhance the scalability and affordability of OR domain modeling and opens a pathway for future advancements in surgical data science. We will release our code and data upon acceptance.

Prior-RadGraphFormer: A Prior-Knowledge-Enhanced Transformer for Generating Radiology Graphs from X-Rays

Y. Xiong, J. Liu, K. Zaripova, S. Sharifzadeh, **M. Keicher***, and N. Navab*, *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology*, 2024 (*Equal contribution.)

The extraction of structured clinical information from free-text radiology reports in the form of radiology graphs has been demonstrated to be a valuable approach for evaluating the clinical correctness of report-generation methods. However, the direct generation of radiology graphs from chest X-ray (CXR) images has not been attempted. To address this gap, we propose a novel approach called Prior-RadGraphFormer that utilizes a transformer model with prior knowledge in the form of a probabilistic knowledge graph (PKG) to generate radiology graphs directly from CXR images. The PKG models the statistical relationship between radiology entities, including anatomical structures and medical observations. This additional contextual information enhances the accuracy of entity and relation extraction. The generated radiology graphs can be applied to various downstream tasks, such as free-text or structured reports generation and multi-label classification of pathologies. Our approach represents a promising method for generating radiology graphs directly from CXR images, and has significant potential for improving medical image analysis and clinical decision-making. Our code is open sourced at <https://github.com/xiongyiheng/Prior-RadGraphFormer>.

Semantic Latent Space Regression of Diffusion Autoencoders for Vertebral Fracture Grading

M. Keicher*, M. Atad*, D. Schinz, A. S. Gersing, S. C. Foreman, S. S. Goller, J. Weissinger, J. Rischewski, A.-S. Dietrich, B. Wiestler, J. S. Kirschke, and N. Navab, *arXiv Preprint 2303.12031*, presented at *iMIMIC 2023 - Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI, 2023* (*Equal contribution. **Best paper award.**)

Vertebral fractures are a consequence of osteoporosis, with significant health implications for affected patients. Unfortunately, grading their severity using CT exams is hard and subjective, motivating automated grading methods. However, current approaches are hindered by imbalance and scarcity of data and a lack of interpretability. To address these challenges, this paper proposes a novel approach that leverages unlabelled data to train a generative Diffusion Autoencoder (DAE) model as an unsupervised feature extractor. We model fracture grading as a continuous regression, which is more reflective of the smooth progression of fractures. Specifically, we use a binary, supervised fracture classifier to construct a hyperplane in the DAE's latent space. We then regress the severity of the fracture as a function of the distance to this hyperplane, calibrating the results to the Genant scale. Importantly, the generative nature of our method allows us to visualize different grades of a given vertebra, providing interpretability and insight into the features that contribute to automated grading.

RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance

C. Pellegrini*, E. Özsoy*, B. Busam, N. Navab, and **M. Keicher**, *arXiv Preprint 2311.18681*, 2023 (*Equal contribution.)

Conversational AI tools that can generate and discuss clinically correct radiology reports for a given medical image have the potential to transform radiology. Such a human-in-the-loop radiology assistant could facilitate a collaborative diagnostic process, thus saving time and improving the quality of reports. Towards this goal, we introduce RaDialog, the first thoroughly evaluated and publicly available large vision-language model for radiology report generation and interactive dialog. RaDialog effectively integrates visual image features and structured pathology findings with a large language model (LLM) while simultaneously adapting it to a specialized domain using parameter-efficient fine-tuning. To keep the conversational abilities of the underlying LLM, we propose a comprehensive, semi-automatically labeled, image-grounded instruct dataset for chest X-ray radiology tasks. By training with this dataset, our method achieves state-of-the-art clinical correctness in report generation and shows impressive abilities in interactive tasks such as correcting reports and answering questions, serving as a foundational step toward clinical dialog systems. Our code is available on github: <https://github.com/ChantalMP/RaDialog>.

Vol2Flow: Segment 3D Volumes Using a Sequence of Registration Flows

A. Bitarafan, M. F. Azampour, K. Bakhtari, M. Soleymani Baghshah, **M. Keicher**, and N. Navab, *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2022.

This work proposes a self-supervised algorithm to segment each arbitrary anatomical structure in a 3D medical image produced under various acquisition conditions, dealing with domain shift problems and generalizability. Furthermore, we advocate an interactive setting in the inference time, where the self-supervised model trained on unlabeled volumes should be directly applicable to segment each test volume given the user-provided single slice annotation. To this end, we learn a novel 3D registration network, namely Vol2Flow, from the perspective of image sequence registration to find 2D displacement fields between all adjacent slices within a 3D medical volume together. Specifically, we present a novel 3D CNN-based architecture that finds a series of registration flows between consecutive slices within a whole volume, resulting in a dense displacement field. A new self-supervised algorithm is proposed to learn the transformations or registration fields between the series of 2D images of a 3D volume. Consequently, we enable gradually propagating the user-provided single slice annotation to other slices of a volume in the inference time. We demonstrate that our model substantially outperforms related methods on various medical image segmentation tasks through several experiments on different medical image segmentation datasets. Code is available at <https://github.com/AdelehBitarafan/Vol2Flow>.

Reproduced with permission from Springer Nature.

Longitudinal Self-Supervision for COVID-19 Pathology Quantification

T. Czempiel, C. Rogers, **M. Keicher**, M. Paschali, R. Braren, E. Burian, M. Makowski, N. Navab, T. Wendler, and S. T. Kim, *arXiv Preprint 2203.10804*, 2022.

Quantifying COVID-19 infection over time is an important task to manage the hospitalization of patients during a global pandemic. Recently, deep learning-based approaches have been proposed to help radiologists automatically quantify COVID-19 pathologies on longitudinal CT scans. However, the learning process of deep learning methods demands extensive training data to learn the complex characteristics of infected regions over longitudinal scans. It is challenging to collect a large-scale dataset, especially for longitudinal training. In this study, we want to address this problem by proposing a new self-supervised learning method to effectively train longitudinal networks for the quantification of COVID-19 infections. For this purpose, longitudinal self-supervision schemes are explored on clinical longitudinal COVID-19 CT scans. Experimental results show that the proposed method is effective, helping the model

better exploit the semantics of longitudinal data and improve two COVID-19 quantification tasks.

U-PET: MRI-based Dementia Detection with Joint Generation of Synthetic FDG-PET Images

M. Kollovieh*, **M. Keicher***, S. Wunderlich, H. Burwinkel, T. Wendler, and N. Navab, *arXiv Preprint 2206.08078*, 2022 (*Equal contribution.)

Alzheimer's disease (AD) is the most common cause of dementia. An early detection is crucial for slowing down the disease and mitigating risks related to the progression. While the combination of MRI and FDG-PET is the best image-based tool for diagnosis, FDG-PET is not always available. The reliable detection of Alzheimer's disease with only MRI could be beneficial, especially in regions where FDG-PET might not be affordable for all patients. To this end, we propose a multi-task method based on U-Net that takes T1-weighted MR images as an input to generate synthetic FDG-PET images and classifies the dementia progression of the patient into cognitive normal (CN), cognitive impairment (MCI), and AD. The attention gates used in both task heads can visualize the most relevant parts of the brain, guiding the examiner and adding interpretability. Results show the successful generation of synthetic FDG-PET images and a performance increase in disease classification over the naive single-task baseline.

ToxNet: an artificial intelligence designed for decision support for toxin prediction

T. Zellner, K. Romanek, C. Rabe, S. Schmoll, S. Geith, E.-C. Heier, R. Stich, H. Burwinkel, **M. Keicher**, D. Bani-Harouni, N. Navab, S.-A. Ahmadi, and F. Eyer, *Clinical Toxicology*, 2022.

BACKGROUND Artificial intelligences (AIs) are emerging in the field of medical informatics in many areas. They are mostly used for diagnosis support in medical imaging but have potential uses in many other fields of medicine where large datasets are available.

AIM To develop an artificial intelligence (AI) "ToxNet", a machine-learning based computer-aided diagnosis (CADx) system, which aims to predict poisons based on patient's symptoms and metadata from our Poison Control Center (PCC) data. To prove its accuracy and compare it against medical doctors (MDs).

METHODS The CADx system was developed and trained using data from 781,278 calls recorded in our PCC database from 2001 to 2019. All cases were mono-intoxications. Patient symptoms and meta-information (e.g., age group, sex, etiology, toxin point of entry, weekday, etc.) were provided. In the pilot phase, the AI was trained on 10 substances, the AI's prediction

was compared to naïve matching, literature matching, a multi-layer perceptron (MLP), and the graph attention network (GAT). The trained AI's accuracy was then compared to 10 medical doctors in an individual and in an identical dataset. The dataset was then expanded to 28 substances and the predictions and comparisons repeated.

RESULTS In the pilot, the prediction performance in a set of 8995 patients with 10 substances was 0.66 ± 0.01 (F1 micro score). Our CADx system was significantly superior to naïve matching, literature matching, MLP, and GAT ($p < 0.005$). It outperformed our physicians experienced in clinical toxicology in the individual and identical dataset. In the extended dataset, our CADx system was able to predict the correct toxin in a set of 36,033 patients with 28 substances with an overall performance of 0.27 ± 0.01 (F1 micro score), also significantly superior to naïve matching, literature matching, MLP, and GAT. It also outperformed our MDs.

CONCLUSION Our AI trained on a large PCC database works well for poison prediction in these experiments. With further research, it might become a valuable aid for physicians in predicting unknown substances and might be the first step into AI use in PCCs.

AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging

D. M. Hedderich, **M. Keicher**, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, T. Czempel, J. E. Spiro, D. Pinto dos Santos, D. Heim, C. Zimmer, D. Rückert, J. S. Kirschke, and N. Navab, *Healthcare*, 2021.

Successful adoption of artificial intelligence (AI) in medical imaging requires medical professionals to understand underlying principles and techniques. However, educational offerings tailored to the need of medical professionals are scarce. To fill this gap, we created the course “AI for Doctors: Medical Imaging”. An analysis of participants’ opinions on AI and self-perceived skills rated on a five-point Likert scale was conducted before and after the course. The participants’ attitude towards AI in medical imaging was very optimistic before and after the course. However, deeper knowledge of AI and the process for validating and deploying it resulted in significantly less overoptimism with respect to perceivable patient benefits through AI ($p = 0.020$). Self-assessed skill ratings significantly improved after the course, and the appreciation of the course content was very positive. However, we observed a substantial drop-out rate, mostly attributed to the lack of time of medical professionals. There is a high demand for educational offerings regarding AI in medical imaging among medical professionals, and better education may lead to a more realistic appreciation of clinical adoption. However, time constraints imposed by a busy clinical schedule need to be taken into account for successful education of medical professionals.

Longitudinal Quantitative Assessment of COVID-19 Infection Progression from Chest CTs

S. T. Kim, L. Goli, M. Paschali, A. Khakzar, **M. Keicher**, T. Czempiel, E. Burian, R. Braren, N. Navab, and T. Wendler, *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2021.

Chest computed tomography (CT) has played an essential diagnostic role in assessing patients with COVID-19 by showing disease-specific image features such as ground-glass opacity and consolidation. Image segmentation methods have proven to help quantify the disease and even help predict the outcome. The availability of longitudinal CT series may also result in an efficient and effective method to reliably assess the progression of COVID-19, monitor the healing process and the response to different therapeutic strategies. In this paper, we propose a new framework to identify infection at a voxel level (identification of healthy lung, consolidation, and ground-glass opacity) and visualize the progression of COVID-19 using sequential low-dose non-contrast CT scans. In particular, we devise a longitudinal segmentation network that utilizes the reference scan information to improve the performance of disease identification. Experimental results on a clinical longitudinal dataset collected in our institution show the effectiveness of the proposed method compared to the static deep neural networks for disease quantification.

Reproduced with permission from Springer Nature.

GLOWin: A Flow-based Invertible Generative Framework for Learning Disentangled Feature Representations in Medical Images

A. Sankar*, **M. Keicher***, R. Eisawy, A. Parida, F. Pfister, S. T. Kim, and N. Navab, *arXiv Preprint 2103.10868*, 2021 (*Equal contribution.)

Disentangled representations can be useful in many downstream tasks, help to make deep learning models more interpretable, and allow for control over features of synthetically generated images that can be useful in training other models that require a large number of labelled or unlabelled data. Recently, flow-based generative models have been proposed to generate realistic images by directly modeling the data distribution with invertible functions. In this work, we propose a new flow-based generative model framework, named GLOWin, that is end-to-end invertible and able to learn disentangled representations. Feature disentanglement is achieved by factorizing the latent space into components such that each component learns the representation for one generative factor. Comprehensive experiments have been conducted to evaluate the proposed method on a public brain tumor MR dataset. Quantitative and qualitative results suggest that the proposed method is effective in disentangling the features from complex medical images.

TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks

T. Czempiel, M. Paschali, **M. Keicher**, W. Simson, H. Feussner, S. T. Kim, and N. Navab, *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2020.

Automatic surgical phase recognition is a challenging and crucial task with the potential to improve patient safety and become an integral part of intra-operative decision-support systems. In this paper, we propose, for the first time in workflow analysis, a Multi-Stage Temporal Convolutional Network (MS-TCN) that performs hierarchical prediction refinement for surgical phase recognition. Causal, dilated convolutions allow for a large receptive field and online inference with smooth predictions even during ambiguous transitions. Our method is thoroughly evaluated on two datasets of laparoscopic cholecystectomy videos with and without the use of additional surgical tool information. Outperforming various state-of-the-art LSTM approaches, we verify the suitability of the proposed causal MS-TCN for surgical phase recognition.

Reproduced with permission from Springer Nature.

Continual Class Incremental Learning for CT Thoracic Segmentation

A. Elskhawy, A. Lisowska, **M. Keicher**, J. Henry, P. Thomson, and N. Navab, *Workshop on Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 2020.

Deep learning organ segmentation approaches require large amounts of annotated training data, which is limited in supply due to reasons of confidentiality and the time required for expert manual annotation. Therefore, being able to train models incrementally without having access to previously used data is desirable. A common form of sequential training is fine tuning (FT). In this setting, a model learns a new task effectively, but loses performance on previously learned tasks. The Learning without Forgetting (LwF) approach addresses this issue via replaying its own prediction for past tasks during model training. In this work, we evaluate FT and LwF for class incremental learning in multi-organ segmentation using the publicly available AAPM dataset. We show that LwF can successfully retain knowledge on previous segmentations, however, its ability to learn a new class decreases with the addition of each class. To address this problem we propose an adversarial continual learning segmentation approach (ACLSeg), which disentangles feature space into task-specific and task-invariant features. This enables preservation of performance on past tasks and effective acquisition of new knowledge.

Reproduced with permission from Springer Nature.

Bibliography

- [1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol. “Multimodal Biomedical AI”. In: *Nature Medicine* 28.9 (Sept. 2022), pp. 1773–1784 (cit. on pp. 7–9, 14).
- [2] N. N. Agu, J. T. Wu, H. Chao, I. Lourentzou, A. Sharma, M. Moradi, P. Yan, and J. Hendler. “Anaxnet: Anatomy aware multi-label finding classification in chest x-ray”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 804–813 (cit. on pp. 82, 83, 85, 87).
- [3] D. M. Allen. “The relationship between variable selection and data augmentation and a method for prediction”. In: *technometrics* 16.1 (1974), pp. 125–127 (cit. on p. 55).
- [4] M. Z. Alom, M. M. S. Rahman, M. S. Nasrin, T. M. Taha, and V. K. Asari. “COVID_MTNNet: COVID-19 Detection with Multi-Task Deep Learning Approaches”. In: *arXiv* (Apr. 2020) (cit. on p. 47).
- [5] A. Amyar, R. Modzelewski, H. Li, and S. Ruan. “Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation”. In: *Computers in Biology and Medicine* 126 (Nov. 2020), p. 104037 (cit. on p. 47).
- [6] R. Anirudh and J. J. Thiagarajan. “Bootstrapping graph convolutional neural networks for autism spectrum disorder classification”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3197–3201 (cit. on p. 30).
- [7] Z. Babar, T. van Laarhoven, and E. Marchiori. “Encoder-Decoder Models for Chest X-ray Report Generation Perform No Better than Unconditioned Baselines”. In: *PLoS ONE* 16.11 (2021). pmid: 34843509 (cit. on p. 73).
- [8] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. “Data2vec: A general framework for self-supervised learning in speech, vision and language”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1298–1312 (cit. on p. 24).
- [9] D. Bahri, H. Jiang, Y. Tay, and D. Metzler. “Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption”. In: *International Conference on Learning Representations*. 2021 (cit. on p. 19).
- [10] G. Bao and X. Wang. “COVID-MTL: Multitask Learning with Shift3D and Random-weighted Loss for Diagnosis and Severity Assessment of COVID-19”. In: *arXiv e-prints* (2020), arXiv–2012 (cit. on p. 47).
- [11] A. Bar, L. Wolf, O. B. Amitai, E. Toledano, and E. Elnekave. “Compression fractures detection on CT”. In: *Medical imaging 2017: computer-aided diagnosis*. Vol. 10134. International Society for Optics and Photonics. 2017, p. 1013440 (cit. on p. 109).
- [12] T. Bartalena, G. Giannelli, M. F. Rinaldi, E. Rimondi, G. Rinaldi, N. Sverzellati, and G. Gavelli. “Prevalence of thoracolumbar vertebral fractures on multidetector CT”. In: *European Journal of Radiology* 69.3 (Mar. 2009), pp. 555–559 (cit. on p. 107).

- [13] O. Bashkanov, M. Rak, L. Engelage, and C. Hansen. “Automatic Patient-Level Diagnosis of Prostate Disease with Fused 3D MRI and Tabular Clinical Data”. In: *Medical Imaging with Deep Learning, MIDL 2023, 10-12 July 2023, Nashville, TN, USA*. Ed. by I. Oguz, J. H. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heimann, D. Kontos, B. A. Landman, and B. M. Dawant. Vol. 227. Proceedings of Machine Learning Research. PMLR, 2023, pp. 1225–1238 (cit. on p. 25).
- [14] R. T. B. Batista-Navarro, D. A. Bandojo, M. A. K. Gatapia, R. N. C. Santos, A. B. Marcelo, L. C. R. Panganiban, and P. C. Naval. “ESP: An expert system for poisoning diagnosis and management”. In: *Informatics for Health and Social Care* 35.2 (2010), pp. 53–63 (cit. on p. 33).
- [15] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. “Network dissection: Quantifying interpretability of deep visual representations”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6541–6549 (cit. on pp. 105, 108–111, 113).
- [16] N. J. Beauchamp, R. N. Bryan, M. M. Bui, G. P. Krestin, G. B. McGinty, C. C. Meltzer, and M. Neumaier. “Integrative Diagnostics: The Time Is Now—a Report from the International Society for Strategic Studies in Radiology”. In: *Insights into Imaging* 14.1 (1 Dec. 2023), pp. 1–13 (cit. on pp. 7–9, 11, 13, 14).
- [17] I. Beltagy, K. Lo, and A. Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620 (cit. on p. 84).
- [18] A. Belyaeva, J. Cosentino, F. Hormozdiari, K. Eswaran, S. Shetty, G. Corrado, A. Carroll, C. Y. McLean, and N. A. Furlotte. “Multimodal LLMs for Health Grounded in Individual-Specific Data”. In: *Machine Learning for Multimodal Healthcare Data*. Ed. by A. K. Maier, J. A. Schnabel, P. Tiwari, and O. Stegle. Cham: Springer Nature Switzerland, 2024, pp. 86–102 (cit. on pp. 23, 121).
- [19] Y. Bengio, A. Courville, and P. Vincent. “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (Aug. 2013), pp. 1798–1828 (cit. on pp. 13, 17).
- [20] W. Berrios, G. Mittal, T. Thrush, D. Kiela, and A. Singh. “Towards language models that can see: Computer vision through the lens of natural language”. In: *arXiv preprint arXiv:2306.16410* (2023) (cit. on p. 22).
- [21] R. Bhalodia, A. Hatamizadeh, L. Tam, Z. Xu, X. Wang, E. Turkbey, and D. Xu. “Improving Pneumonia Localization via Cross-Attention on Medical Images and Reports”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 571–581 (cit. on p. 77).
- [22] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer. “Retrieval-Augmented Diffusion Models”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 15309–15324 (cit. on p. 30).
- [23] D. Bliuc. “Mortality Risk Associated With Low-Trauma Osteoporotic Fracture and Subsequent Fracture in Men and Women”. en. In: *JAMA* 301.5 (Feb. 2009), p. 513 (cit. on p. 107).
- [24] A. Blomberg, M. Keicher, and F. Weber. “Die Möglichkeiten zur Durchführung Systematischer Früherkennungsuntersuchungen bei Krebserkrankungen”. In: *DGOR: Papers of the 10th Annual Meeting/Vorträge der 10. Jahrestagung*. Springer. 1982, pp. 171–177 (cit. on p. 6).
- [25] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay. “Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing”. In: *Computer Vision – ECCV 2022*. Ed. by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner. Cham: Springer Nature Switzerland, 2022, pp. 1–21 (cit. on pp. 17, 24, 25, 78, 91, 92, 94).
- [26] L. Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 19).

- [27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 24).
- [28] L. Buess, M. F. Stollenga, D. Schinz, B. Wiestler, J. Kirschke, A. Maier, N. Navab, and M. Keicher. “Video-CT MAE: Self-supervised Video-CT Domain Adaptation for Vertebral Fracture Diagnosis”. In: *Accepted for presentation at Medical Imaging with Deep Learning*. 2024 (cit. on p. 18).
- [29] E. Burian, F. Jungmann, G. A. Kaissis, et al. “Intensive Care Risk Estimation in COVID-19 Pneumonia Based on Clinical and Imaging Parameters: Experiences from the Munich Cohort”. In: *Journal of Clinical Medicine* 9 (5 May 2020), p. 1514 (cit. on pp. 21, 44, 46, 50, 51, 55, 58).
- [30] H. Burwinkel, A. Kazi, G. Vivar, S. Albarqouni, G. Zahnd, N. Navab, and S.-A. Ahmadi. “Adaptive Image-Feature Learning for Disease Classification Using Inductive Graph Networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 640–648 (cit. on pp. 30, 47).
- [31] H. Burwinkel, M. Keicher, D. Bani-Harouni, T. Zellner, F. Eyer, N. Navab, and S.-A. Ahmadi. “Decision support for intoxication prediction using graph convolutional networks”. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II* 23. Springer. 2020, pp. 633–642 (cit. on pp. 33, 35, 38–40).
- [32] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang. “A multimodal transformer to fuse images and metadata for skin disease classification”. In: *The Visual Computer* (2022), pp. 1–13 (cit. on p. 45).
- [33] W. Cai, T. Liu, X. Xue, G. Luo, X. Wang, Y. Shen, Q. Fang, J. Sheng, F. Chen, and T. Liang. “CT Quantification and Machine-learning Models for Assessment of Disease Severity and Prognosis of COVID-19 Patients”. In: *Academic Radiology* 27 (12 Dec. 2020), pp. 1665–1678 (cit. on p. 46).
- [34] Y.-H. Cao, H. Yu, and J. Wu. “Training vision transformers with only 2040 images”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 220–237 (cit. on p. 18).
- [35] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660 (cit. on p. 24).
- [36] J. A. Cauley. “Public Health Impact of Osteoporosis”. en. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 68.10 (Oct. 2013), pp. 1243–1251 (cit. on p. 107).
- [37] J. Cauley, D. Thompson, K. Ensrud, J. Scott, and D. Black. “Risk of mortality following clinical fractures”. In: *Osteoporosis international* 11.7 (2000), pp. 556–561 (cit. on p. 107).
- [38] J. R. Center, T. V. Nguyen, D. Schneider, P. N. Sambrook, and J. A. Eisman. “Mortality after all major types of osteoporotic fracture in men and women: an observational study”. en. In: *The Lancet* 353.9156 (Mar. 1999), pp. 878–882 (cit. on p. 107).
- [39] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari. “Roentgen: Vision-language foundation model for chest x-ray generation”. In: *arXiv preprint arXiv:2211.12737* (2022) (cit. on p. 23).
- [40] H. Chao, X. Fang, J. Zhang, et al. “Integrative analysis for COVID-19 patient outcome prediction”. In: *Medical Image Analysis* 67 (Jan. 2021), p. 101844 (cit. on pp. 46, 58).
- [41] G. Chassagnon, M. Vakalopoulou, E. Battistella, et al. “AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia”. In: *Medical Image Analysis* 67 (Jan. 2021), p. 101860 (cit. on p. 46).
- [42] *ChatGPT by OpenAI*. chat.openai.com. Accessed: 2023-03-08 (cit. on p. 94).

- [43] G. Chauhan, R. Liao, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, and P. Golland. “Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 529–539 (cit. on p. 77).
- [44] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang. “Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification”. In: *IEEE Journal of Biomedical and Health Informatics* 24.8 (2020), pp. 2292–2302 (cit. on p. 77).
- [45] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. “This Looks Like That: Deep Learning for Interpretable Image Recognition”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 29).
- [46] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison. *Deep Hierarchical Multi-label Classification of Chest X-ray Images*. Ed. by M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren. July 2019 (cit. on p. 77).
- [47] H.-Y. Chen, B. W.-Y. Hsu, Y.-K. Yin, F.-H. Lin, T.-H. Yang, R.-S. Yang, C.-K. Lee, and V. S. Tseng. “Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs”. In: *Plos one* 16.1 (2021), e0245992 (cit. on p. 109).
- [48] T. Chen and C. Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cit. on p. 19).
- [49] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang. “Multi-modal masked autoencoders for medical vision-and-language pre-training”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 679–689 (cit. on p. 24).
- [50] D. Chetrit, T. Meir, H. Lebel, M. Orlovsky, R. Gordon, A. Akselrod-Ballin, and A. Bar. “3D convolutional sequence to sequence model for vertebral compression fractures identification in CT”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 743–752 (cit. on p. 109).
- [51] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. “3D U-Net: learning dense volumetric segmentation from sparse annotation”. In: *MICCAI*. Springer. 2016, pp. 424–432 (cit. on p. 110).
- [52] D. Colombi, F. C. Bodini, M. Petrini, G. Maffi, N. Morelli, G. Milanese, M. Silva, N. Sverzellati, and E. Michieletti. “Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia”. In: *Radiology* 296 (2 Aug. 2020), E86–E96 (cit. on p. 46).
- [53] Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, and The National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. Ed. by E. P. Balogh, B. T. Miller, and J. R. Ball. Pages: 21794. Washington, D.C.: National Academies Press, Dec. 2015 (cit. on pp. 6–8).
- [54] L. Cosmo, A. Kazi, S.-A. Ahmadi, N. Navab, and M. Bronstein. “Latent-graph learning for disease prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 643–653 (cit. on p. 30).
- [55] P. Croskerry. “A Universal Model of Diagnostic Reasoning”. In: *Academic Medicine* 84.8 (Aug. 2009), p. 1022 (cit. on p. 9).
- [56] T. Czempiel, M. Paschali, **M. Keicher**, W. Simson, H. Feussner, S. T. Kim, and N. Navab. “TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks”. en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 343–352 (cit. on p. 19).

- [57] S. J. Darmoni, P. Massari, J. M. Droy, N. Mahe, T. Blanc, E. Moiro, and J. Leroy. “SETH: an expert system for the management on acute drug poisoning in adults”. In: *Computer Methods and Programs in Biomedicine* 43.3-4 (1994), pp. 171–176 (cit. on p. 33).
- [58] M. Defferrard, X. Bresson, and P. Vandergheynst. “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. In: *Advances in Neural Information Processing Systems* (2016), pp. 3844–3852 (cit. on p. 33).
- [59] L. Degenhardt, F. Charlson, A. Ferrari, et al. “The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016”. In: *The Lancet Psychiatry* 5.12 (2018), pp. 987–1012 (cit. on p. 33).
- [60] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. “Preparing a Collection of Radiology Examinations for Distribution and Retrieval”. In: *Journal of the American Medical Informatics Association : JAMIA* 23.2 (Mar. 2016), pp. 304–310. pmid: 26133894 (cit. on p. 16).
- [61] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*. NAACL-HLT 2019. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186 (cit. on pp. 17, 22, 24).
- [62] D. Di, F. Shi, F. Yan, et al. “Hypergraph learning for identification of COVID-19 with CT imaging”. In: *Medical Image Analysis* 68 (2021), p. 101910 (cit. on p. 46).
- [63] A. Di Martino, C.-G. Yan, Q. Li, et al. “The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism”. In: *Molecular Psychiatry* 19.6 (June 2014), pp. 659–667 (cit. on p. 16).
- [64] B. Djulbegovic and G. H. Guyatt. “Progress in Evidence-Based Medicine: A Quarter Century On”. In: *The Lancet* 390.10092 (July 2017), pp. 415–423 (cit. on pp. 6, 10).
- [65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2021 (cit. on pp. 18, 22, 84).
- [66] D. P. dos Santos, E. Kotter, P. Mildemberger, L. Martí-Bonmatí, and European Society of Radiology (ESR). “ESR Paper on Structured Reporting in Radiology—Update 2023”. In: *Insights into Imaging* 14.1 (Nov. 23, 2023), p. 199 (cit. on p. 72).
- [67] H. Duanmu, P. B. Huang, S. Brahmavar, S. Lin, T. Ren, J. Kong, F. Wang, and T. Q. Duong. “Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data”. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II* 23. Springer. 2020, pp. 242–252 (cit. on p. 45).
- [68] D. E. Dupuy, A. E. Rosenberg, T. Punyaratabandhu, M. H. Tan, and H. J. Mankin. “Accuracy of CT-Guided Needle Biopsy of Musculoskeletal Neoplasms”. In: *American Journal of Roentgenology* 171.3 (Sept. 1998), pp. 759–762 (cit. on p. 107).
- [69] M. Endo, R. Krishnan, V. Krishna, A. Y. Ng, and P. Rajpurkar. “Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model”. In: *Proceedings of Machine Learning for Health*. Machine Learning for Health. PMLR, Nov. 28, 2021, pp. 209–219 (cit. on pp. 30, 122).

- [70] M. Endo, R. Krishnan, V. Krishna, A. Y. Ng, and P. Rajpurkar. “Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model”. In: *Proceedings of Machine Learning for Health*. Ed. by S. Roy, S. Pfohl, E. Rocheteau, G. A. Tadesse, L. Oala, F. Falck, Y. Zhou, L. Shen, G. Zamzmi, P. Mugambi, A. Zirikly, M. B. A. McDermott, and E. Alsentzer. Vol. 158. Proceedings of Machine Learning Research. PMLR, Dec. 2021, pp. 209–219 (cit. on p. 25).
- [71] P. Engstler, M. Keicher, D. Schinz, K. Mach, A. S. Gersing, S. C. Foreman, S. S. Goller, J. Weissinger, J. Rischewski, A.-S. Dietrich, B. Wiestler, J. S. Kirschke, A. Khakzar, and N. Navab. “Interpretable Vertebral Fracture Diagnosis”. en. In: *Interpretability of Machine Intelligence in Medical Image Computing*. Ed. by M. Reyes, P. Henriques Abreu, and J. Cardoso. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 71–81 (cit. on pp. 108, 113–115).
- [72] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. “Deep Learning-Enabled Medical Computer Vision”. In: *npj Digital Medicine* 4.1 (Jan. 8, 2021), pp. 1–9 (cit. on p. 17).
- [73] S. Feng, B. Liu, Y. Zhang, X. Zhang, and Y. Li. “Two-Stream Compare and Contrast Network for Vertebral Compression Fracture Diagnosis”. In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pp. 2496–2506 (cit. on p. 109).
- [74] M. Fey and J. E. Lenssen. “Fast graph representation learning with PyTorch Geometric”. In: *arXiv preprint arXiv:1903.02428* (2019) (cit. on p. 57).
- [75] O. Fink, Q. Wang, M. Svensen, P. Dersin, W.-J. Lee, and M. Ducoffe. “Potential, challenges and future directions for deep learning in prognostics and health management applications”. In: *Engineering Applications of Artificial Intelligence* 92 (2020), p. 103678 (cit. on p. 91).
- [76] D. Ganeshan, P.-A. T. Duong, L. Probyn, L. Lenchik, T. A. McArthur, M. Retrouvey, E. H. Ghobadi, S. L. Desouches, D. Pastel, and I. R. Francis. “Structured Reporting in Radiology”. In: *Academic Radiology* 25.1 (Jan. 2018), pp. 66–73. PMID: 29030284 (cit. on p. 72).
- [77] K. Gao, J. Su, Z. Jiang, L. L. Zeng, Z. Feng, H. Shen, P. Rong, X. Xu, J. Qin, Y. Yang, W. Wang, and D. Hu. “Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images”. In: *Medical Image Analysis* 67 (Jan. 2021), p. 101836 (cit. on p. 47).
- [78] H. K. Genant, C. Y. Wu, C. Van Kuijk, and M. C. Nevitt. “Vertebral fracture assessment using a semiquantitative technique”. In: *Journal of bone and mineral research* 8.9 (1993), pp. 1137–1148 (cit. on p. 113).
- [79] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. “Imagebind: One embedding space to bind them all”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15180–15190 (cit. on p. 122).
- [80] O. Goksel, A. Foncubierta-Rodríguez, O. A. J. del Toro, et al. “Overview of the VISCERAL Challenge at ISBI 2015”. In: *VISCERAL Challenge@ISBI*. 2015 (cit. on p. 55).
- [81] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals”. In: *circulation* 101.23 (2000), e215–e220 (cit. on p. 81).
- [82] M. Goncharov, M. Pisov, A. Shevtsov, B. Shirokikh, A. Kurmukov, I. Blokhin, V. Chernina, A. Solovev, V. Gombolevskiy, S. Morozov, et al. “Ct-based covid-19 triage: Deep multitask learning improves joint identification and severity quantification”. In: *Medical image analysis* 71 (2021), p. 102054 (cit. on pp. 47, 49).
- [83] K. Gong, D. Wu, C. D. Arru, F. Homayounieh, N. Neumark, J. Guan, V. Buch, K. Kim, B. C. Bizzo, H. Ren, et al. “A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records”. In: *European journal of radiology* 139 (2021), p. 109583 (cit. on p. 46).

- [84] Y. Gorishniy, I. Rubachev, V. Khruikov, and A. Babenko. “Revisiting deep learning models for tabular data”. In: *Advances in Neural Information Processing Systems 34* (2021), pp. 18932–18943 (cit. on p. 19).
- [85] J. Haczynski and A. Jakimiuk. “Vertebral fractures: a hidden problem of osteoporosis”. eng. In: *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research 7.5* (Oct. 2001), pp. 1108–1117 (cit. on p. 107).
- [86] P. Hager, F. Jungmann, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, R. Holland, R. Braren, M. Makowski, and G. Kaisis. “Evaluating and Mitigating Limitations of Large Language Models in Clinical Decision Making”. In: *medRxiv* (2024), pp. 2024–01 (cit. on p. 121).
- [87] P. Hager, M. J. Menten, and D. Rueckert. “Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, June 2023, pp. 23924–23935 (cit. on p. 25).
- [88] I. Hallberg, M. Bachrach-Lindström, S. Hammerby, G. Toss, and A.-C. Ek. “Health-related quality of life after vertebral or hip fracture: a seven-year follow-up study”. en. In: *BMC Musculoskeletal Disorders 10.1* (Dec. 2009), p. 135 (cit. on p. 107).
- [89] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, B. Wittmann, E. Simsar, M. Simsar, et al. “A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities”. In: *arXiv preprint arXiv:2403.17834* (2024) (cit. on pp. 15, 16).
- [90] W. L. Hamilton, R. Ying, and J. Leskovec. “Inductive representation learning on large graphs”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 1025–1035 (cit. on p. 58).
- [91] C. R. Harris, K. J. Millman, S. J. van der Walt, et al. “Array programming with NumPy”. In: *Nature 585.7825* (Sept. 2020), pp. 357–362 (cit. on p. 57).
- [92] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on pp. 17, 18, 58).
- [93] K. He, W. Zhao, X. Xie, W. Ji, M. Liu, Z. Tang, Y. Shi, F. Shi, Y. Gao, J. Liu, et al. “Synergistic learning of lung lobe segmentation and hierarchical multi-instance classification for automated severity assessment of COVID-19 in CT images”. In: *Pattern recognition 113* (2021), p. 107828 (cit. on p. 47).
- [94] L. Heiliger, A. Sekuboyina, B. Menze, J. Egger, and J. Kleesiek. “Beyond medical imaging-A review of multimodal deep learning in radiology”. In: *Authorea Preprints* (2023) (cit. on p. 23).
- [95] E. Hernlund, A. Svedbom, M. Ivergård, J. Compston, C. Cooper, J. Stenmark, E. V. McCloskey, B. Jönsson, and J. A. Kanis. “Osteoporosis in the European Union: medical management, epidemiology and economic burden: A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA)”. en. In: *Archives of Osteoporosis 8.1-2* (Dec. 2013), p. 136 (cit. on p. 107).
- [96] L. S. Hesse and A. I. L. Namburete. “INSightR-Net: Interpretable Neural Network for Regression Using Similarity-Based Comparisons to Prototypical Examples”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li. Cham: Springer Nature Switzerland, 2022, pp. 502–511 (cit. on p. 30).
- [97] J. Higgs, G. M. Jensen, S. Loftus, and N. Christensen, eds. *Clinical Reasoning in the Health Professions*. Fourth edition. Edinburgh London New York: Elsevier, 2019. 511 pp. (cit. on pp. 10, 11).

- [98] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs. “Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem”. In: *European Radiology Experimental* 4.1 (2020), p. 50 (cit. on p. 55).
- [99] F. Homayounieh, R. Babaei, H. K. Mobin, C. D. Arru, M. Sharifian, I. Mohseni, E. Zhang, S. R. Digumarthy, and M. K. Kalra. “Computed tomography radiomics can predict disease severity and outcome in coronavirus disease 2019 pneumonia”. In: *Journal of Computer Assisted Tomography* 44 (5 2020), pp. 640–646 (cit. on p. 46).
- [100] Y. Hong and C. E. Kahn. “Content analysis of reporting templates and free-text radiology reports”. In: *Journal of digital imaging* 26.5 (2013), pp. 843–849 (cit. on p. 75).
- [101] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz. “RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 293–303 (cit. on p. 77).
- [102] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708 (cit. on pp. 17, 84, 110).
- [103] H. Huang, M. Cai, L. Lin, J. Zheng, X. Mao, X. Qian, Z. Peng, J. Zhou, Y. Iwamoto, X.-H. Han, et al. “Graph-based Pyramid Global Context Reasoning with a Saliency-aware Projection for COVID-19 Lung Infections Segmentation”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 1050–1054 (cit. on p. 46).
- [104] S.-C. Huang, Z. Huo, E. Steinberg, C.-C. Chiang, C. Langlotz, M. Lungren, S. Yeung, N. Shah, and J. Fries. “INSPECT: A Multimodal Dataset for Patient Outcome Prediction of Pulmonary Embolisms”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 17742–17772 (cit. on pp. 15, 16).
- [105] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. “Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines”. In: *npj Digital Medicine* 3.1 (1 Oct. 16, 2020), pp. 1–9 (cit. on pp. 13, 21, 22).
- [106] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. “Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3942–3951 (cit. on pp. 77, 78, 91).
- [107] M. Hussein, A. Sekuboyina, M. Loeffler, F. Navarro, B. H. Menze, and J. S. Kirschke. “Grading loss: a fracture grade-based metric loss for vertebral fracture detection”. In: *MICCAI*. Springer. 2020 (cit. on pp. 109, 112, 113).
- [108] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597 (cit. on pp. 15, 16, 73, 92, 94, 100).
- [109] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18.2 (2021), pp. 203–211 (cit. on pp. 18, 55).
- [110] S. Jain, A. Agrawal, A. Saporta, S. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, C. Langlotz, and P. Rajpurkar. “RadGraph: Extracting Clinical Entities and Relations from Radiology Reports”. en. In: *PhysioNet*. June 2021 (cit. on pp. 72, 89).
- [111] T. Jalava, S. Sarna, L. Pylkkänen, B. Mawer, J. A. Kanis, P. Selby, M. Davies, J. Adams, R. M. Francis, J. Robinson, and E. McCloskey. “Association Between Vertebral Fracture and Increased Mortality in Osteoporotic Patients”. en. In: *Journal of Bone and Mineral Research* 18.7 (July 2003), pp. 1254–1260 (cit. on p. 107).

- [112] R. Jana and I. Mukherjee. “Deep Learning Based Efficient Epileptic Seizure Prediction with EEG Channel Optimization”. In: *Biomedical Signal Processing and Control* 68 (July 1, 2021), p. 102767 (cit. on p. 19).
- [113] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. “DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-language in Genome”. In: *Bioinformatics* 37.15 (Aug. 9, 2021), pp. 2112–2120 (cit. on p. 19).
- [114] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, and Y. Zhu. “Few-shot Radiology Report Generation for Rare Diseases”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 601–608 (cit. on p. 78).
- [115] E. Jimenez-Solem, T. S. Petersen, C. Hansen, et al. “Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients”. In: *Scientific Reports* 11.1 (2021) (cit. on p. 46).
- [116] Y. Jin, Y. Chen, L. Wang, J. Wang, P. Yu, L. Liang, J.-N. Hwang, and Z. Liu. “Decoupling Object Detection from Human-Object Interaction Recognition”. In: *arXiv preprint arXiv:2112.06392* (2021) (cit. on pp. 79, 80).
- [117] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. “MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0)”. In: *PhysioNet* (2019) (cit. on pp. 15, 16, 81).
- [118] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042* (2019) (cit. on pp. 81, 100).
- [119] D. Kahneman. *Thinking, Fast and Slow*. 1st pbk. ed. New York: Farrar, Straus and Giroux, 2013 (cit. on p. 9).
- [120] **M. Keicher**, M. Atad, D. Schinz, A. S. Gersing, S. C. Foreman, S. S. Goller, J. Weissinger, J. Rischewski, A.-S. Dietrich, B. Wiestler, J. S. Kirschke, and N. Navab. *Semantic Latent Space Regression of Diffusion Autoencoders for Vertebral Fracture Grading*. arXiv:2303.12031 [cs]. Mar. 2023 (cit. on pp. 24, 105).
- [121] **M. Keicher**, H. Burwinkel, D. Bani-Harouni, M. Paschali, T. Czempiel, E. Burian, M. R. Makowski, R. Braren, N. Navab, and T. Wendler. “Multimodal graph attention network for COVID-19 outcome prediction”. en. In: *Scientific Reports* 13.1 (Nov. 2023). Number: 1 Publisher: Nature Publishing Group, p. 19539 (cit. on pp. 44–46, 48, 52–54, 60–65, 110).
- [122] **M. Keicher**, K. Zaripova, T. Czempiel, K. Mach, A. Khakzar, and N. Navab. “FlexR: Few-shot Classification with Language Embeddings for Structured Reporting of Chest X-rays”. en. In: Nashville: Proceedings of Machine Learning Research, Apr. 2023 (cit. on pp. 76, 81, 82, 85, 87, 91).
- [123] A. Khakzar, S. Musatian, J. Buchberger, I. Valeriano Quiroz, N. Pinger, S. Baselizadeh, S. T. Kim, and N. Navab. “Towards Semantic Interpretation of Thoracic Disease and COVID-19 Diagnosis Models”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert. Cham: Springer International Publishing, 2021, pp. 499–508 (cit. on pp. 105, 108–110).
- [124] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar. “Mmbert: Multimodal bert pretraining for improved medical vqa”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1033–1036 (cit. on p. 24).
- [125] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine. “Supervised multimodal bitransformers for classifying images and text”. In: *arXiv preprint arXiv:1909.02950* (2019) (cit. on p. 22).
- [126] D. Kiela, E. Grave, A. Joulin, and T. Mikolov. “Efficient large-scale multi-modal classification”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018 (cit. on p. 22).

- [127] E. Kim, S. Kim, M. Seo, and S. Yoon. “XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15719–15728 (cit. on p. 30).
- [128] S. T. Kim, L. Goli, M. Paschali, A. Khakzar, **M. Keicher**, T. Czempiel, E. Burian, R. Braren, N. Navab, and T. Wendler. “Longitudinal quantitative assessment of COVID-19 infection progression from chest CTs”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer. 2021, pp. 273–282 (cit. on pp. 19, 52).
- [129] T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2016 (cit. on p. 33).
- [130] M. Kohli, T. Alkasab, K. Wang, M. E. Heilbrun, A. E. Flanders, K. Dreyer, and C. E. Kahn. “Bending the Artificial Intelligence Curve for Radiology: Informatics Tools From ACR and RSNA”. In: *Journal of the American College of Radiology*. Focus Issue: Survey Methods 16.10 (Oct. 1, 2019), pp. 1464–1470 (cit. on p. 72).
- [131] M. Kollovieh, **M. Keicher**, S. Wunderlich, H. Burwinkel, T. Wendler, and N. Navab. *U-PET: MRI-based Dementia Detection with Joint Generation of Synthetic FDG-PET Images*. arXiv:2206.08078 [cs, eess]. June 2022 (cit. on p. 23).
- [132] E. Kotter and D. Pinto dos Santos. “Strukturierte Befundung in der Radiologie”. In: *Der Radiologe* 61.11 (Nov. 1, 2021), pp. 979–985 (cit. on p. 72).
- [133] C. Koufidis, K. Manninen, J. Nieminen, M. Wohlin, and C. Silén. “Unravelling the Polyphony in Clinical Reasoning Research in Medical Education”. In: *Journal of Evaluation in Clinical Practice* 27.2 (2021), pp. 438–450 (cit. on pp. 9, 10).
- [134] P. Kulling and H. Persson. “Role of the intensive care unit in the management of the poisoned patient”. In: *Medical toxicology* 1.5 (1986), pp. 375–86 (cit. on p. 33).
- [135] M. Larose, N. Touma, N. Raymond, D. LeBlanc, F. Rasekh, B. Neveu, H. Hovington, M. Vallières, F. Pouliot, and L. Archambault. “Graph Attention Network for Prostate Cancer Lymph Node Invasion Prediction”. In: *Medical Imaging with Deep Learning*. 2022 (cit. on p. 49).
- [136] D. B. Larson and C. P. Langlotz. “The Role of Radiology in the Diagnostic Process: Information, Communication, and Teamwork”. In: *American Journal of Roentgenology* 209.5 (Nov. 2017), pp. 992–1000 (cit. on p. 8).
- [137] T.-L.-T. Le, N. Thome, S. Bernard, V. Bismuth, and F. Patoureaux. “Multitask classification and segmentation for cancer diagnosis in mammography”. In: *arXiv preprint arXiv:1909.05397* (2019) (cit. on p. 46).
- [138] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. “Temporal convolutional networks: A unified approach to action segmentation”. In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 47–54 (cit. on p. 19).
- [139] Y. LeCun, Y. Bengio, and G. Hinton. “Deep Learning”. In: *Nature* 521.7553 (May 2015), pp. 436–444 (cit. on p. 17).
- [140] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474 (cit. on pp. 30, 122).
- [141] J. Li, C. Liu, S. Cheng, R. Arcucci, and S. Hong. “Frozen Language Model Helps ECG Zero-Shot Learning”. In: *Medical Imaging with Deep Learning, MIDL 2023, 10–12 July 2023, Nashville, TN, USA*. Ed. by I. Oguz, J. H. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heimann, D. Kontos, B. A. Landman, and B. M. Dawant. Vol. 227. *Proceedings of Machine Learning Research*. PMLR, 2023, pp. 402–415 (cit. on p. 25).

- [142] J. Li, D. Li, S. Savarese, and S. Hoi. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742 (cit. on p. 24).
- [143] Y. Li, Y. Zhang, E. Zhang, Y. Chen, Q. Wang, K. Liu, H. J. Yu, H. Yuan, N. Lang, and M.-Y. Su. “Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning”. In: *European Radiology* 31.12 (2021), pp. 9612–9619 (cit. on p. 109).
- [144] X. Liang, Y. Zhang, J. Wang, Q. Ye, Y. Liu, and J. Tong. “Diagnosis of COVID-19 Pneumonia Based on Graph Convolutional Network”. In: *Frontiers in Medicine* 7 (2021), p. 1071 (cit. on p. 46).
- [145] R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, and W. M. Wells. “Multimodal Representation Learning via Maximization of Local Mutual Information”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 273–283 (cit. on p. 77).
- [146] J. Lipkova, R. J. Chen, B. Chen, M. Y. Lu, M. Barbieri, D. Shao, A. J. Vaidya, C. Chen, L. Zhuang, D. F. Williamson, M. Shaban, T. Y. Chen, and F. Mahmood. “Artificial Intelligence for Multimodal Data Integration in Oncology”. In: *Cancer Cell* 40.10 (Oct. 2022), pp. 1095–1110 (cit. on p. 22).
- [147] H. Liu, C. Li, Q. Wu, and Y. J. Lee. “Visual instruction tuning”. In: *Advances in neural information processing systems* 36 (2024) (cit. on p. 23).
- [148] X. Liu, H. Wang, Z. Li, and L. Qin. “Deep Learning in ECG Diagnosis: A Review”. In: *Knowledge-Based Systems* 227 (Sept. 5, 2021), p. 107187 (cit. on p. 19).
- [149] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel. “Retrieval Augmented Classification for Long-Tail Visual Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6959–6969 (cit. on p. 30).
- [150] J. B. Long, Y. Zhang, V. Brusica, L. Chitkushev, and G. Zhang. “Antidote Application”. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM Press, 2017, pp. 442–448 (cit. on p. 33).
- [151] J. Lu, D. Batra, D. Parikh, and S. Lee. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 24).
- [152] S. M. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777 (cit. on p. 110).
- [153] L. Luo, Y.-F. Li, G. Haffari, and S. Pan. “Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning”. In: *International Conference on Learning Representations*. 2024 (cit. on p. 122).
- [154] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. “BioGPT: generative pre-trained transformer for biomedical text generation and mining”. In: *Briefings in bioinformatics* 23.6 (2022), bbac409 (cit. on p. 17).
- [155] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C, G. Ramakrishnan, M. Goldblum, and C. White. “When do neural nets outperform boosted trees on tabular data?” In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 19, 20).
- [156] D. McInerney, G. Young, J.-W. van de Meent, and B. Wallace. “CHILL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8477–8494 (cit. on p. 91).

- [157] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro. “Y-Net: joint segmentation and classification for diagnosis of breast biopsy images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 893–901 (cit. on p. 46).
- [158] L. J. Melton III, E. J. Atkinson, C. Cooper, W. M. O’Fallon, and B. L. Riggs. “Vertebral Fractures Predict Subsequent Fractures”. In: *Osteoporosis International* 10.3 (1999) (cit. on p. 107).
- [159] Y. Meng, M. Wei, D. Gao, Y. Zhao, X. Yang, X. Huang, and Y. Zheng. “CNN-GCN aggregation enabled boundary regression for biomedical image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 352–362 (cit. on p. 30).
- [160] S. Menon and C. Vondrick. “Visual Classification via Description from Large Language Models”. In: *arXiv preprint arXiv:2210.07183* (2022) (cit. on p. 91).
- [161] C. Messiou, R. Lee, and M. Salto-Tellez. “Multimodal Analysis and the Oncology Patient: Creating a Hospital System for Integrated Diagnostics and Discovery”. In: *Computational and Structural Biotechnology Journal* 21 (Jan. 1, 2023), pp. 4536–4539 (cit. on pp. 7–9).
- [162] A. N. D. Meyer, T. D. Giardina, L. Khawaja, and H. Singh. “Patient and Clinician Experiences of Uncertainty in the Diagnostic Process: Current Understanding and Future Directions”. In: *Patient Education and Counseling* 104.11 (Nov. 1, 2021), pp. 2606–2615 (cit. on p. 7).
- [163] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571 (cit. on pp. 18, 56).
- [164] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller. “Explaining nonlinear classification decisions with deep Taylor decomposition”. In: *Pattern Recognition* (2017). arXiv: 1512.02479 (cit. on p. 110).
- [165] S. M. Monteiro and G. Norman. “Diagnostic Reasoning: Where We’ve Been, Where We’re Going”. In: *Teaching and Learning in Medicine* 25.sup1 (Jan. 2013), S26–S32 (cit. on pp. 10, 11).
- [166] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. “Foundation Models for Generalist Medical Artificial Intelligence”. In: *Nature* 616.7956 (7956 Apr. 2023), pp. 259–265 (cit. on p. 12).
- [167] K. Murata, K. Endo, T. Aihara, H. Suzuki, Y. Sawaji, Y. Matsuoka, H. Nishimura, T. Takamatsu, T. Konishi, A. Maekawa, et al. “Artificial intelligence for the detection of vertebral fractures on plain spinal radiography”. In: *Scientific Reports* (2020) (cit. on p. 109).
- [168] J. J. Näppi, T. Uemura, C. Watari, T. Hironaka, T. Kamiya, and H. Yoshida. “U-survival for prognostic prediction of disease progression and mortality of patients with COVID-19”. In: *Scientific reports* 11.1 (2021), pp. 1–11 (cit. on p. 47).
- [169] J. Nicolaes, S. Raeymaeckers, D. Robben, G. Wilms, D. Vandermeulen, C. Libanati, and M. Debois. “Detection of vertebral fractures in CT using 3D convolutional neural networks”. In: *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*. Springer. 2019, pp. 3–14 (cit. on p. 109).
- [170] W. Ning, S. Lei, J. Yang, et al. “Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning”. In: *Nature Biomedical Engineering* 4 (12 Dec. 2020), pp. 1197–1207 (cit. on pp. 16, 46, 55, 56).
- [171] J. M. Nobel, E. M. Kok, and S. G. Robben. “Redefining the structure of structured reporting in radiology”. In: *Insights into Imaging* 11.1 (2020), pp. 1–5 (cit. on p. 75).
- [172] J. M. Nobel, K. van Geel, and S. G. F. Robben. “Structured reporting in radiology: a systematic review to explore its potential”. In: *European Radiology* (Oct. 2021) (cit. on p. 75).

- [173] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. “Capabilities of gpt-4 on medical challenge problems”. In: *arXiv preprint arXiv:2303.13375* (2023) (cit. on p. 121).
- [174] G. Norman. “Research in clinical reasoning: past history and current trends”. en. In: *Medical Education* 39.4 (Apr. 2005), pp. 418–427 (cit. on p. 6).
- [175] G. R. Norman, S. D. Monteiro, J. Sherbino, J. S. Ilgen, H. G. Schmidt, and S. Mamede. “The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking”. In: *Academic Medicine* 92.1 (Jan. 2017), pp. 23–30 (cit. on p. 10).
- [176] E. Özsoy, C. Pellegrini, **M. Keicher**, and N. Navab. “ORacle: Large Vision-Language Models for Knowledge-Guided Holistic OR Domain Modeling”. In: *arXiv Preprint 2404.07031*. 2024 (cit. on p. 121).
- [177] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert. “Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease”. In: *Medical image analysis* 48 (2018), pp. 117–130 (cit. on pp. 30, 44).
- [178] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, and D. Rueckert. “Spectral Graph Convolutions for Population-based Disease Prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Verlag, 2017, pp. 177–185 (cit. on p. 34).
- [179] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, and D. Rueckert. “Spectral graph convolutions for population-based disease prediction”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 177–185 (cit. on pp. 30, 57).
- [180] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019), pp. 8026–8037 (cit. on p. 57).
- [181] V. L. Patel, D. R. Kaufman, and J. F. Arocha. “Emerging paradigms of cognition in medical decision-making”. In: *Journal of biomedical informatics* 35.1 (2002), pp. 52–75 (cit. on p. 12).
- [182] A. Paul, Y.-X. Tang, T. Shen, and R. Summers. “Discriminative ensemble learning for few-shot chest x-ray diagnosis”. In: *Medical Image Analysis* 68 (Feb. 2021), p. 101911 (cit. on p. 78).
- [183] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 57–59).
- [184] C. Pellegrini, **M. Keicher**, E. Özsoy, P. Jiraskova, R. Braren, and N. Navab. “Xplainer: From X-Ray Observations to Explainable Zero-Shot Diagnosis”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Lecture Notes in Computer Science. Springer Nature Switzerland, Oct. 2023, pp. 420–429 (cit. on pp. 91, 93, 95–100).
- [185] C. Pellegrini, **M. Keicher**, E. Özsoy, and N. Navab. “Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting”. en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 409–419 (cit. on pp. 23, 72, 73).
- [186] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and **M. Keicher**. *RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance*. arXiv:2311.18681 [cs]. Nov. 2023 (cit. on pp. 20, 22, 23, 73, 121).
- [187] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. “FiLM: Visual Reasoning with a General Conditioning Layer”. 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. New Orleans, Louisiana, USA: AAAI Press, 2018 (cit. on p. 45).

- [188] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. “Alzheimer’s Disease Neuroimaging Initiative (ADNI)”. In: *Neurology* 74.3 (Jan. 19, 2010), pp. 201–209. pmid: 20042704 (cit. on p. 16).
- [189] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen. “Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels”. In: *Neurocomputing* 437 (2021), pp. 186–194 (cit. on p. 77).
- [190] P. Pino, D. Parra, C. Besa, and C. Lagos. “Clinically Correct Report Generation from Chest X-Rays Using Templates”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, 2021, pp. 654–663 (cit. on pp. 72, 73, 76, 77).
- [191] M. Pisov, V. Kondratenko, A. Zakharov, A. Petraikin, V. Gombolevskiy, S. Morozov, and M. Belyaev. “Keypoints localization for joint vertebra detection and fracture severity quantification”. In: *MICCAI*. Springer, 2020, pp. 723–732 (cit. on p. 109).
- [192] C. Qin, D. Yao, Y. Shi, and Z. Song. “Computer-aided detection in chest radiography based on artificial intelligence: a survey”. In: *Biomedical engineering online* 17.1 (2018), pp. 1–23 (cit. on p. 91).
- [193] R. Raab, A. Küderle, A. Zakreuskaya, A. D. Stern, J. Klucken, G. Kaissis, D. Rueckert, S. Boll, R. Eils, H. Wagener, and B. M. Eskofier. “Federated Electronic Health Records for the European Health Data Space”. In: *The Lancet Digital Health* 5.11 (Nov. 1, 2023), e840–e847. pmid: 37741765 (cit. on pp. 71, 122).
- [194] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763 (cit. on pp. 24, 25, 80, 84, 91, 94).
- [195] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. “Improving language understanding by generative pre-training”. In: (2018) (cit. on p. 15).
- [196] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. “AI in Health and Medicine”. In: *Nature Medicine* 28.1 (1 Jan. 2022), pp. 31–38 (cit. on pp. 11, 17).
- [197] D. Ramachandram and G. W. Taylor. “Deep Multimodal Learning: A Survey on Recent Advances and Trends”. In: *IEEE Signal Processing Magazine* 34.6 (Nov. 2017), pp. 96–108 (cit. on pp. 13, 22).
- [198] A. Remuzzi and G. Remuzzi. “COVID-19 and Italy: what next?” eng. In: *Lancet (London, England)* 395.10231 (Apr. 2020), pp. 1225–1228 (cit. on p. 43).
- [199] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241 (cit. on pp. 18, 49, 56).
- [200] B. C. Ross. “Mutual information between discrete and continuous data sets”. In: *PloS one* 9.2 (2014), e87357 (cit. on pp. 48, 49, 57).
- [201] H. R. Roth, Z. Xu, C. T. Diez, R. S. Jacob, J. Zember, J. Molto, W. Li, S. Xu, B. Turkbey, E. Turkbey, et al. “Rapid artificial intelligence solutions in a pandemic—the COVID-19-20 lung CT lesion segmentation challenge”. In: *Research Square* (2021) (cit. on p. 55).
- [202] R. D. Rudyanto, S. Kerkstra, E. M. van Rikxoort, et al. “Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study”. In: *Medical image analysis* 18 7 (2014), pp. 1217–32 (cit. on p. 55).
- [203] J. Ryberg. “COVID-19, triage decisions, and indirect ethics: A model for the re-evaluation of triage guidelines”. In: *Ethics, Medicine and Public Health* 17 (2021), p. 100639 (cit. on p. 43).

- [204] P. Saha, D. Mukherjee, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar. “GraphCovidNet: A graph neural network based model for detecting COVID-19 from CT scans and X-rays of chest”. In: *Scientific Reports* 11.1 (2021), pp. 1–16 (cit. on p. 46).
- [205] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin. “Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods”. In: *Computers in Biology and Medicine* 140 (Jan. 1, 2022), p. 105111 (cit. on pp. 12, 105).
- [206] L. Santhosh, C. L. Chou, and D. M. Connor. “Diagnostic Uncertainty: From Education to Communication”. In: *Diagnosis* 6.2 (June 1, 2019), pp. 121–126 (cit. on p. 7).
- [207] I. H. Sarker. “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions”. In: *SN Computer Science* 2.6 (Aug. 18, 2021), p. 420 (cit. on p. 19).
- [208] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf. “Restricting the Flow: Information Bottlenecks for Attribution”. In: *International Conference on Learning Representations*. 2020 (cit. on p. 110).
- [209] C. Seibold, S. Reiß, M. S. Sarfraz, R. Stiefelhagen, and J. Kleesiek. “Breaking With Fixed Set Pathology Recognition Through Report-Guided Contrastive Training”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*. Singapore, Singapore: Springer-Verlag, 2022, pp. 690–700 (cit. on pp. 78, 88, 91, 94, 96).
- [210] M. Seibold, A. Hoch, M. Farshad, N. Navab, and P. Fürnstahl. “Conditional generative data augmentation for clinical audio datasets”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 345–354 (cit. on p. 19).
- [211] A. Sekuboyina, M. E. Hussein, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern, et al. “VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images”. In: *Medical image analysis* (2021) (cit. on pp. 108, 111).
- [212] H. S. Shin. “Reasoning Processes in Clinical Reasoning: From the Perspective of Cognitive Psychology”. In: *Korean Journal of Medical Education* 31.4 (Dec. 2019), pp. 299–308. PMID: 31813196 (cit. on pp. 5, 9–11).
- [213] I. Shiri, M. Sorouri, P. Geramifar, M. Nazari, M. Abdollahi, Y. Salimi, B. Khosravi, D. Askari, L. Aghaghazvini, G. Hajianfar, et al. “Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients”. In: *Computers in biology and medicine* 132 (2021), p. 104304 (cit. on p. 46).
- [214] R. Shwartz-Ziv and A. Armon. “Tabular Data: Deep Learning Is Not All You Need”. In: *Information Fusion* 81 (May 1, 2022), pp. 84–90 (cit. on p. 19).
- [215] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. “CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT”. In: *arXiv preprint arXiv:2004.09167* (2020) (cit. on p. 73).
- [216] R. D. Soberanis-Mukul, N. Navab, and S. Albarqouni. “Uncertainty-based graph convolutional networks for organ segmentation refinement”. In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 755–769 (cit. on p. 30).
- [217] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein. “Saint: Improved neural networks for tabular data via row attention and contrastive pre-training”. In: *arXiv preprint arXiv:2106.01342* (2021) (cit. on p. 19).
- [218] X. Song, H. Li, W. Gao, Y. Chen, T. Wang, G. Ma, and B. Lei. “Augmented Multi-center Graph Convolutional Network for COVID-19 Diagnosis”. In: *IEEE Transactions on Industrial Informatics* (2021) (cit. on p. 46).
- [219] H. C. Sox, M. C. Higgins, and D. K. Owens. *Medical decision making*. 2nd ed. Chichester, West Sussex, UK : Hoboken, New Jersey: John Wiley & Sons, 2013 (cit. on pp. 6–8).

- [220] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. “Emu: Generative Pretraining in Multimodality”. In: *The Twelfth International Conference on Learning Representations*. Oct. 13, 2023 (cit. on p. 122).
- [221] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. “An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success”. In: *npj Digital Medicine* 3.1 (1 Feb. 6, 2020), pp. 1–10 (cit. on p. 11).
- [222] T. Syeda-Mahmood, K. C. Wong, Y. Gur, J. T. Wu, A. Jadhav, S. Kashyap, A. Karargyris, A. Pillai, A. Sharma, A. B. Syed, et al. “Chest X-ray report generation through fine-grained label learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 561–571 (cit. on pp. 72, 77).
- [223] A. Taleb, M. Kirchler, R. Monti, and C. Lippert. “Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20908–20921 (cit. on p. 25).
- [224] A. Taleb, C. Lippert, T. Klein, and M. Nabi. “Multimodal self-supervised learning for medical image analysis”. In: *International conference on information processing in medical imaging*. Springer. 2021, pp. 661–673 (cit. on p. 24).
- [225] Y. Tang, H. Yang, L. Zhang, and Y. Yuan. “Work like a Doctor: Unifying Scan Localizer and Dynamic Generator for Automated Computed Tomography Report Generation”. In: *Expert Systems with Applications* 237 (Mar. 1, 2024), p. 121442 (cit. on pp. 15, 16).
- [226] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, T. Ma, J. Liu, and D. Shen. “Severity assessment of COVID-19 using CT image features and laboratory indices”. In: *Physics in Medicine & Biology* 66 (3 Feb. 2021) (cit. on p. 46).
- [227] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. “Any-to-any generation via composable diffusion”. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on pp. 24, 122).
- [228] A. Tariq, L. A. Celi, J. M. Newsome, S. Purkayastha, N. K. Bhatia, H. Trivedi, J. W. Gichoya, and I. Banerjee. “Patient-specific COVID-19 resource utilization prediction using fusion AI model”. In: *NPJ digital medicine* 4.1 (2021), pp. 1–9 (cit. on p. 46).
- [229] Z. Tian, X. Li, Y. Zheng, Z. Chen, Z. Shi, L. Liu, and B. Fei. “Graph-convolutional-network-based interactive prostate segmentation in MR images”. In: *Medical physics* 47.9 (2020), pp. 4164–4176 (cit. on p. 30).
- [230] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning”. In: *Nature Biomedical Engineering* (2022), pp. 1–8 (cit. on pp. 78, 88, 91, 94, 96).
- [231] N. Tomita, Y. Y. Cheung, and S. Hassanpour. “Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans”. In: *Computers in biology and medicine* 98 (2018), pp. 8–15 (cit. on p. 109).
- [232] A. Valentinitich, S. Trebeschi, J. Kaesmacher, C. Lorenz, M. Löffler, C. Zimmer, T. Baum, and J. Kirschke. “Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures”. In: *Osteoporosis international* 30.6 (2019), pp. 1275–1285 (cit. on p. 109).
- [233] D. Van Veen, C. Van Uden, L. Blankemeier, et al. “Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization”. In: *Nature Medicine* (Feb. 27, 2024), pp. 1–9 (cit. on p. 121).
- [234] S. van Baalen, M. Boon, and P. Verhoef. “From Clinical Decision Support to Clinical Reasoning Support Systems”. In: *Journal of Evaluation in Clinical Practice* 27.3 (2021), pp. 520–528 (cit. on pp. 11, 12).

- [235] H. Vaseli, A. N. Gu, S. N. Ahmadi Amiri, M. Y. Tsang, A. Fung, N. Kondori, A. Saadat, P. Abolmaesumi, and T. S. M. Tsang. “ProtoASNet: Dynamic Prototypes for Inherently Interpretable and Uncertainty-Aware Aortic Stenosis Classification in Echocardiography”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Cham: Springer Nature Switzerland, 2023, pp. 368–378 (cit. on p. 30).
- [236] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 15, 22).
- [237] P. Veličković, A. Casanova, P. Liò, G. Cucurull, A. Romero, and Y. Bengio. “Graph attention networks”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2018 (cit. on pp. 34, 36, 50).
- [238] P. Virtanen, R. Gommers, T. E. Oliphant, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272 (cit. on p. 57).
- [239] D. Wang, C. Huang, S. Bao, T. Fan, Z. Sun, Y. Wang, H. Jiang, and S. Wang. “Study on the prognosis predictive model of COVID-19 patients based on CT radiomics”. In: *Scientific reports* 11.1 (2021), pp. 1–9 (cit. on p. 46).
- [240] S.-H. Wang, V. V. Govindaraj, J. M. Górriz, X. Zhang, and Y.-D. Zhang. “Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network”. In: *Information Fusion* 67 (2021), pp. 208–229 (cit. on p. 46).
- [241] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, et al. “Image as a foreign language: Beit pretraining for all vision and vision-language tasks”. In: *arXiv preprint arXiv:2208.10442* (2022) (cit. on p. 24).
- [242] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106 (cit. on pp. 92, 94).
- [243] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. “Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9049–9058 (cit. on p. 77).
- [244] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. “MedCLIP: Contrastive Learning from Unpaired Medical Images and Text”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Abu Dhabi, United Arab Emirates, 2022, pp. 3876–3887 (cit. on pp. 25, 91).
- [245] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun. “Transformers in Time Series: A Survey”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. Thirty-Second International Joint Conference on Artificial Intelligence {IJCAI-23}. Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 6778–6786 (cit. on p. 19).
- [246] A. L. Williams, A. Al-Busaidi, P. J. Sparrow, J. E. Adams, and R. W. Whitehouse. “Under-reporting of osteoporotic vertebral fractures on computed tomography”. en. In: *European Journal of Radiology* 69.1 (Jan. 2009), pp. 179–183 (cit. on p. 107).
- [247] T. N. Wolf, F. Bongratz, A.-M. Rickmann, S. Pölsterl, and C. Wachinger. “Keep the Faith: Faithful Explanations in Convolutional Neural Networks for Case-Based Reasoning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 6. 2024, pp. 5921–5929 (cit. on p. 30).
- [248] T. N. Wolf, S. Pölsterl, C. Wachinger, A. D. N. Initiative, et al. “DAFT: A universal module to interweave tabular data and 3D images in CNNs”. In: *NeuroImage* 260 (2022), p. 119505 (cit. on p. 45).

- [249] J. M. Wolterink, T. Leiner, and I. Išgum. “Graph convolutional networks for coronary artery segmentation in cardiac CT angiography”. In: *International Workshop on Graph Learning in Medical Imaging*. Springer. 2019, pp. 62–69 (cit. on p. 30).
- [250] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. “MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, Oct. 1, 2023, pp. 21315–21326 (cit. on p. 25).
- [251] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, and M.-M. Cheng. “Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 3113–3126 (cit. on p. 47).
- [252] J. Wu, B. Zhou, D. Peck, S. Hsieh, V. Dialani, L. Mackey, and G. Patterson. “Deepminer: Discovering interpretable representations for mammogram classification and explanation”. In: *arXiv preprint arXiv:1805.12323* (2018) (cit. on pp. 108, 110).
- [253] J. Wu, N. Agu, I. Lourentzou, A. Sharma, J. Paguio, J. S. Yao, E. C. Dee, W. Mitchell, S. Kashyap, A. Giovannini, L. A. Celi, T. Syeda-Mahmood, and M. Moradi. “Chest ImaGenome Dataset (version 1.0.0)”. In: *PhysioNet* (2021) (cit. on pp. 72, 81, 82, 89).
- [254] Y. Xiong, J. Liu, K. Zaripova, S. Sharifzadeh, **M. Keicher**, and N. Navab. “Prior-RadGraphFormer: A Prior-Knowledge-Enhanced Transformer for Generating Radiology Graphs from X-Rays”. In: *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology*. Ed. by S.-A. Ahmadi and S. Pereira. Cham: Springer Nature Switzerland, 2024, pp. 54–63 (cit. on p. 72).
- [255] P. Xu, X. Zhu, and D. A. Clifton. “Multimodal Learning With Transformers: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (Oct. 2023), pp. 12113–12132 (cit. on p. 22).
- [256] Q. Xu, X. Zhan, Z. Zhou, Y. Li, P. Xie, S. Zhang, X. Li, Y. Yu, C. Zhou, L. Zhang, O. Gevaert, and G. Lu. “CT-based Rapid Triage of COVID-19 Patients: Risk Prediction and Progression Estimation of ICU Admission, Mechanical Ventilation, and Death of Hospitalized Patients.” In: *medRxiv : the preprint server for health sciences* (Nov. 2020), p. 2020.11.04.20225797 (cit. on p. 46).
- [257] R. Y. “Differentiation of Osteoporotic and Neoplastic Vertebral Fractures by Chemical Shift {In-Phase and Out-of Phase} Magnetic Resonance Imaging and Diffusion Weighted Sequence”. In: *MOJ Orthopedics & Rheumatology* 6.2 (Nov. 2016) (cit. on p. 107).
- [258] A. Yan, Z. He, X. Lu, J. Du, E. Chang, A. Gentili, J. McAuley, and C.-N. Hsu. “Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation”. In: *arXiv preprint arXiv:2109.12242* (2021) (cit. on p. 77).
- [259] J. Yang, H. Veeraraghavan, S. G. Armato, et al. “Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017”. In: *Medical Physics* 45 (2018), pp. 4568–4581 (cit. on p. 55).
- [260] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su. “A novel multi-task deep learning model for skin lesion segmentation and classification”. In: *arXiv preprint arXiv:1703.01025* (2017) (cit. on p. 46).
- [261] E. B. Yilmaz, C. Buerger, T. Fricke, M. M. R. Sagar, J. Peña, C. Lorenz, C.-C. Glüer, and C. Meyer. “Automated Deep Learning-Based Detection of Osteoporotic Fractures in CT Images”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2021 (cit. on p. 109).
- [262] E. B. Yilmaz, A. O. Mader, T. Fricke, J. Peña, C.-C. Glüer, and C. Meyer. “Assessing attribution maps for explaining CNN-based vertebral fracture classifiers”. In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020 (cit. on p. 109).
- [263] W. J. Youden. “Index for rating diagnostic tests”. In: *Cancer* 3.1 (1950), pp. 32–35 (cit. on p. 59).

- [264] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, et al. “Evaluating progress in automatic chest x-ray radiology report generation”. In: *Patterns* 4.9 (2023) (cit. on p. 73).
- [265] X. Yu, S. Lu, L. Guo, S.-H. Wang, and Y.-D. Zhang. “ResGNet-C: A graph convolutional neural network for detection of COVID-19”. In: *Neurocomputing* (2020) (cit. on p. 46).
- [266] T. Zellner, K. Romanek, C. Rabe, S. Schmoll, S. Geith, E.-C. Heier, R. Stich, H. Burwinkel, **M. Keicher**, D. Bani-Harouni, N. Navab, S.-A. Ahmadi, and F. Eyer. “ToxNet: an artificial intelligence designed for decision support for toxin prediction”. In: *Clinical Toxicology* 61.1 (Jan. 2023). Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/15563650.2022.2144345>, pp. 56–63 (cit. on p. 33).
- [267] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. “BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs”. In: *arXiv preprint arXiv:2303.00915* (2023) (cit. on p. 25).
- [268] Y. Zhang, A. Khakzar, Y. Li, A. Farshad, S. T. Kim, and N. Navab. “Fine-Grained Neural Network Explanation by Identifying Input Features with Predictive Information”. In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 110).
- [269] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu. “When radiology report generation meets knowledge graph”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 2020, pp. 12910–12917 (cit. on p. 77).
- [270] Y. Zhang, P. Tiño, A. Leonardis, and K. Tang. “A Survey on Neural Network Interpretability”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (Oct. 2021), pp. 726–742 (cit. on p. 105).
- [271] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. “Contrastive Learning of Medical Visual Representations from Paired Images and Text”. In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. Ed. by Z. Lipton, R. Ranganath, M. Sendak, M. Sjoding, and S. Yeung. Vol. 182. Proceedings of Machine Learning Research. PMLR, Aug. 2022, pp. 2–25 (cit. on pp. 25, 91).
- [272] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929 (cit. on pp. 108, 110).
- [273] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, et al. “Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr”. In: *arXiv preprint arXiv:2111.11665* (2021) (cit. on p. 16).

List of Figures

1.1	Overview of the contributions of this thesis.	4
2.1	The clinical decision-making process. The process begins when a patient presents with a chief concern. The clinician generates a differential diagnosis based on the patient’s symptoms and available data. Next, the differential is narrowed to the most likely diagnosis through an iterative clinical reasoning process. This diagnosis is then communicated to the patient, and a treatment plan is established. The clinician relies on formal and experiential knowledge throughout this process, continuously improving through practice and education. This process is adapted from the diagnostic process outlined by Balogh et al. [53] incorporating the cyclic reasoning process of differential diagnosis described by Sox et al. [219].	6
2.2	Integrated diagnostics. To find the best treatment for a patient, clinicians need to form a holistic representation of the patient to confirm or rule out hypotheses in the differential diagnosis. Adapted from [1, 16, 53, 161]	7
2.3	Uncertainty reduction during the differential diagnosis process. Adapted from Sox et al. [219]	8
3.1	ResNet50 is a well-known Convolutional Neural Network commonly used for encoding and classifying medical images proposed by He et al. [92]. Here, the classification of 14 findings is visualized with a synthetic X-ray.	18
3.2	There are three primary types of multimodal deep learning: First, the fusion of modalities (Section 3.3.1) to solve a downstream task, such as classifying findings in a chest X-ray by using both the image and the clinical history written as text. Second, the translation (Section 3.3.2) from one modality to another, like generating a report for a given chest X-ray image. Third, learning multimodal representations (Section 3.3.3) without explicit supervision, such as using contrastive learning.	20
3.3	There are three main ways to fuse different modalities of data: early fusion, joint fusion, and late fusion.	21
5.1	The patient symptom vectors are processed in parallel by a population Graph Convolutional Network (GCN) and a symptom-matching network. The population GCN, based on Graph Attention Network (GAT) layers, considers both reported symptoms and patient demographics to refine patient representations by considering features of similar patients. Simultaneously, the symptom-matching branch maps patient-reported symptoms to textbook descriptions, integrating formal knowledge into the model. The resulting outputs from both branches are then combined to classify the toxin. [31] <i>Reproduced with permission from Springer Nature.</i>	35

5.2	<p>Left: Evaluation of ToxNet and various baseline approaches using 10-fold cross-validation. Right: Assessment of ToxNet and baseline methods compared to the performance of MDs across 10 distinct datasets, each assessed by a single MD. [31] <i>Reproduced with permission from Springer Nature.</i></p>	39
5.3	<p>Inter-variability among clinicians and comparison to ToxNet’s performance. The toxin categories are arranged in alphabetical order, with each group separated by a small gap. [31] <i>Reproduced with permission from Springer Nature.</i></p>	40
6.1	<p>U-GAT is a multitasking model that segments pathologies and uses this information to predict various patient outcomes. It integrates image features (Z_I), radiomic features (R), and clinical metadata (X_C) and is end-to-end optimized to extract the most relevant features for disease prognosis. The clinical metadata includes information such as age, sex, vital signs, and blood levels. The model segments disease-affected areas (Y_{Seg}) in CT images (X_I), which helps in extracting radiomic features (R) and regularizes the extraction of image features (Z_I). These features are fused into a multimodal vector representation using a function Ψ. Test patients are clustered with training patients in a graph based on the similarity between their radiomic and clinical data features defined by ω. A Graph Attention Network (GAT) then refines the features to predict the most probable outcome (Y) using a learned linear transformation Θ and patient attention coefficients α_{ij}. Outcome classification is supervised using binary cross-entropy (BCE) loss, while the Dice loss is used for the auxiliary segmentation task. Applied to COVID-19, U-GAT segments pathologies in the lung CT image and predicts outcomes such as ICU admission, ventilation needs, survival, or severity. [121]</p>	45
6.2	<p>Visualization of the patient population graph for ICU, ventilation, and survival prediction tasks on the KRI dataset. In the top row, each node is connected to its seven nearest neighbors, determined by calculating the Euclidean distance between feature vectors formed by both clinical and radiomic features. We introduce feature weighting in the distance calculation to optimize the graph structure for a specific task without using prior knowledge about the disease. This weighting is based on the mutual information [200] between features and the task at hand, as shown in the bottom row. This gives features relevant to the task a higher priority in forming patient neighborhoods, encouraging aggregation of features from relevant patients. [121]</p>	48
6.3	<p>Boxplot visualization of the Dice scores for the segmentation results of different methods on the KRI dataset. The multitasking approaches are compared with single-task segmentation. Although the auxiliary segmentation task improves classification results, the segmentation performance is lower in all multitask settings compared to the U-Net optimized solely for segmentation. [121]</p>	61

6.4	KRI dataset - Ensemble results for the ICU task on the KRI dataset: receiver operating characteristic (ROC) and precision-recall curves comparing inner cross-validation loop ensembles of Random Forest and the proposed U-GAT method. Each narrow curve represents the results of one of the 5 test sets. The average of these curves is estimated with the bold curve, and the shaded area depicts its standard deviation. The confusion matrices on the right include the predictions for all patients across all 5 test folds with optimized thresholds. The corresponding metrics of these confusion matrices are visualized with filled circles in the curve diagrams.	63
6.5	Visualization of the neighborhood attention of GAT for a single test patient from the KRI dataset. Left: Batch graph showing the attention scores of the test patient’s neighbors after two hops. The line thickness corresponds to the attention score of each neighbor. Right: CT images, segmentation ground truth, and predicted segmentation of a single axial and coronal slice from the test patient and its neighbor with the maximum attention score. Bottom: Most important features for the test patient and the neighbor with maximum attention, with the radiomics predicted by the pretrained U-Net shown in brackets. [121] .	65
8.1	The proposed Few-shot classification with Language Embeddings for chest X-ray Reporting (FlexR) method builds on self-supervised pre-training to accurately predict fine-grained radiological findings, requiring only a limited number of high-quality annotations. The approach consists of three main stages: (1) contrastive language-image pre-training on a dataset containing radiology images paired with their corresponding unstructured reports, (2) encoding of radiological findings extracted from structured reports, and (3) fine-tuning of the resulting language embeddings to optimize the specific structured reporting task. This process enables the efficient extraction of clinical knowledge from unlabelled, unstructured patient data. [122]	76
8.2	The class distribution in the MIMIC-CXR / ImaGenome task, which involves localizing nine pathologies across 29 anatomical locations, exhibits a long-tailed pattern. The classes, arranged in order of increasing frequency of occurrence and displayed on a logarithmic scale, demonstrate a significant imbalance, with a few classes appearing much more frequently than other classes in the dataset and severely underrepresented classes on the left. [122]	81
8.3	Extraction of structured findings from ImaGenome[253]: Structured reporting elements resembling common data elements (CDEs) are extracted from the ImaGenome knowledge graph by first extracting triplets and then rephrasing them as sentences. [122]	82
8.4	RadReport data elements for severity of cardiomegaly: Structured reports can be represented as a hierarchy of findings often organized by organs of interest. .	83
9.1	Xplainer overview: Observation probabilities are calculated using contrastive CLIP prompting and then utilized to make an explainable diagnosis prediction, providing insights into the model’s decision-making process. [184] <i>Reproduced with permission from Springer Nature.</i>	93

9.2	Qualitative results of Xplainer, demonstrating the interpretability and plausibility of the classification-by-description approach. The model's predictions, even when incorrect, provide valuable insights into its decision-making process, facilitating error detection and interpretation by radiologists. [184] <i>Reproduced with permission from Springer Nature.</i>	100
11.1	Visualization of the detector units most highly correlated with a true positive prediction and clinical experts' interpretation of their activations. All displayed samples are fractured and represented by a slice with high activation after thresholding. [71] <i>Reproduced with permission from Springer Nature.</i>	114
11.2	Visualization of the most relevant detector units during the classification of the sample shown on the left, which the network correctly identifies as fractured. Each detector unit is represented by a single slice activation for that specific sample, and its ranking is based on its high correlation with true positive predictions. The visualization reveals that the network uses concepts associated with wedge-shaped deformity and incorporates information from an adjacent vertebra. [71] <i>Reproduced with permission from Springer Nature.</i>	115

List of Tables

3.1	Overview of data modalities, data types, and common deep learning architectures used in clinical decision support.	14
3.2	Overview of multimodal datasets for various clinical applications.	16
5.1	Evaluation of different methods for toxin prediction. The methods are detailed in Section 5.3 (p-value: <0.01 *, <0.005 **). [31] <i>Reproduced with permission from Springer Nature.</i>	38
6.1	An overview of feature extraction backbones and classifiers used for evaluation with the according modalities used as patient features and similarity metric. <i>Images</i> refer to the image features extracted with an image encoder. <i>Radiomics</i> are the radiomics calculated on the predicted segmentation. <i>Clinical</i> data encompasses vital signs, blood test results, and demographics. U-GAT is compared to a set of end-to-end methods using only clinical data (MLP-Clinical), images (ResNet18), and a U-GAT variant without auxiliary segmentation but a simple ResNet18 instead (ResNet18-GAT). Additionally, we perform experiments on image embeddings extracted from a frozen U-Net, pretrained on the same segmentation task, denoted as U-Net*. Multitasking refers to the joint optimization of classification and segmentation. [121]	52
6.2	KRI dataset - Blood test results upon hospital admission for the 53 ICU patients and the 79 non-ICU patients. The total number of patients (n) is less than the overall study population ($n = 132$) due to missing data for certain individuals. Statistical significance at the 5% level is indicated by *. [121]	53
6.3	KRI dataset - Blood test results at hospital admission for the 38 patients requiring ventilation and the 94 patients who did not. The total n is less than the overall study population ($n = 132$) owing to missing data for some patients. Significant differences at the 5% level are denoted by *. [121]	53
6.4	KRI dataset - Blood test results upon hospital admission for the 113 surviving patients and the 19 deceased patients. The sum of n is lower than the total study population ($n = 132$) due to incomplete data for certain individuals. Statistically significant differences at the 5% level are indicated by *. [121]	54
6.5	KRI dataset -Radiomic features extracted from manually segmented admission CT scans. Statistical significance at the 5% level is denoted by *. [121]	54
6.6	KRI dataset - Top 10 features ranked by their mutual information and Pearson correlation for each task, calculated as the average across the training sets of all repetitions. For the multilabel setup, the mutual information between each feature and the ordinal regression of outcome severity is estimated. [121]	60

6.7	iCTCF dataset - Top 10 features ranked by their mutual information with the outcome severity (Type I vs. Type II) and Pearson correlation, averaged across the training sets of all ten folds. The overall mutual information and Pearson correlation values are lower than the tasks in our in-house dataset. The COVID-19 burden radiomic feature, extracted from the U-Net and equivalent to one minus the healthy lung percentage, consistently ranks among the most important features. [121]	61
6.8	Comparison of edge features and their weighting schemes for distance calculation, evaluated on the validation set of the KRI dataset. [121]	61
6.9	Ablative testing and comparison of the proposed method with an MLP using only clinical data and a ResNet18 using only image data as input for all tasks. U-GAT* denotes the proposed method using image and radiomic features extracted from a frozen U-Net trained on the same annotations as the end-to-end U-GAT. Values marked with † indicate statistical significance ($p < 0.05$) based on Wilcoxon's rank test comparing the proposed method with each baseline. [121]	62
6.10	iCTCF Dataset - Comparison of test set DICE scores between U-Net and U-GAT on the iCTCF dataset. The joint optimization of segmentation and classification leads to a minor decrease in segmentation metrics. [121]	63
6.11	KRI dataset - Results for the multitasking of pathology segmentation and the prediction of three patient outcomes (ICU admission, need for ventilation, and mortality) on the KRI dataset. The graph construction is based on the ordinal regression of outcome severity. Each outcome prediction is modeled as a non-exclusive binary classification, i.e., a multilabel problem. The mortality task is the only task benefiting from this multitask setup. [121]	64
6.12	Comparative analysis of ICU outcome prediction on the KRI dataset: U-GAT versus its cross-validation ensemble, a random forest model using only clinical data, and another random forest model incorporating all available tabular data, including radiomics extracted with a pretrained U-Net. [121]	64
8.1	Grading for cardiomegaly severity in the <i>Chest Xray - 2 Views</i> RadReport template with the label support extracted from MIMIC-CXR reports. [122]	85
8.2	Ablation study of the FlexR method using different backbones and weight initializations for pathology localization, along with results for cardiomegaly grading compared to naïve transfer learning. The evaluation metric is mean AUC. N-shot indicates the number of annotated samples per class used for training. The proposed approach is highlighted in bold . [122]	87
8.3	Comparison of FlexR against baselines using all available data, with and without pathology localization, as well as naïve transfer learning in the few-shot setting. The evaluation metric is AUC, and N-shot refers to the number of annotated samples per class used for training. The proposed method is marked in bold [122].	87
9.1	Descriptors for each pathology used in the Xplainer framework, showcasing the detailed radiological findings that contribute to the diagnosis prediction and explainability of the model. [184] <i>Reproduced with permission from Springer Nature.</i>	95

9.2	AUC scores for zero-shot pathology classification on CheXpert and ChestX-ray14 datasets, comparing different prompting approaches. Results marked with * indicate in-domain testing, as the underlying CLIP model was also trained on the ChestX-ray14 dataset, while the other results report out-of-domain performance. [184] <i>Reproduced with permission from Springer Nature.</i>	96
9.3	AUC per chest X-ray class of clinical findings on the CheXpert validation and test set as well as the ChestX-ray14 test set. [184] <i>Reproduced with permission from Springer Nature.</i>	97
9.4	Comparison of different prompting styles on the CheXpert validation set, demonstrating the effectiveness of contrastive observation-based prompting with pathology indication and report style. [184] <i>Reproduced with permission from Springer Nature.</i>	98
9.5	Comparison of ChatGPT-generated prompts and prompts refined with the help of a senior radiologist, showing the benefit of incorporating domain knowledge into prompt engineering. [184] <i>Reproduced with permission from Springer Nature.</i>	98
9.6	Comparison of modeling the "No Finding" label using explicit prompts or a rule-based definition as the absence of other findings, demonstrating the effectiveness of the rule-based approach. [184] <i>Reproduced with permission from Springer Nature.</i>	98
9.7	Comparison of single-view inference and different methods for multi-view aggregation, highlighting the advantage of averaging observation probabilities across multiple views. [184] <i>Reproduced with permission from Springer Nature.</i>	99
11.1	Evaluation of the trained neural networks' performance on the test holdout of the VerSe dataset and the combined dataset, which includes VerSe and proprietary data obtained from Klinikum rechts der Isar and Klinikum der Universität München. The VerSe dataset consists of 3,920 non-cervical vertebrae, with 254 fractures, while the combined dataset encompasses 10,675 T1-L5 vertebrae, including 1,246 fractures. [71] <i>Reproduced with permission from Springer Nature.</i>	113

