To the Graduate Council:
I am submitting herewith a dissertation written by Christoph S. Metzner entitled "Improving Performance of Clinical Text Classification Models with Attention Mechanisms and Rationales." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Data Science and Engineering.

Heidi A. Hanson, Major Professor

We have read this dissertation
and recommend its acceptance:

Shang Gao

Drahomira Herrmannova

Russell Zaretzki

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Improving Performance of Clinical Text Classification Models with Attention Mechanisms and Rationales

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Christoph S. Metzner

December 2024

Auf den Schultern all meiner Riesen.

On the shoulders of all my giants.

Acknowledgements

This dissertation was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing under Award Number DE-SC-ERKJ422 and by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725.

I would like to express my sincere gratitude to Dr. Heidi Hanson for her continuous support and guidance during the creation of my dissertation. In particular, her ability to provide me with great feedback if warranted while giving me significant autonomy over the direction of my research. I thoroughly enjoyed working with Dr. Hanson and experienced my time with the Oak Ridge National Laboratory as a true blessing.

I would also like to express my sincere gratitude to Dr. Shang Gao and Dr. Drahomira Herrmannova who have supported me since Fall 2021 and were an integral part during countless discussion on my research progress. After a short stint at another group at ORNL in late 2021, Dr. Gao and Dr. Herrmannova went above and beyond to ensure I found an excellent placement to continue my research with

the Biostatistics and Multiscale System Modeling Group and Dr. Heidi Hanson at ORNL.

I am deeply grateful to Dr. Russell Zartezki for serving as the final member of my doctoral committee. Beyond his function as the final committee member, Dr. Zaretzki has shown continuous support throughout my entire Ph.D. journey. During an initially difficult time, Russell encouraged me to take a leave of absence instead of quitting the PhD. He and the Bredesen Center ensured that I was welcomed back and supported.

I would also like to express my sincere gratitude to my previous advisors Dr. Tim M. Young of the University of Tennessee, Dr. Alexander Petuttschnigg and Dr. Marius-Catalin Barbu of the University of Applied Sciences Salzburg in Austria. Only through their ongoing support for students, I was able to come to the United States to participate in a double-degree program. My work with Dr. Young, sparked my passion for programming and statistics and changed my life from being a wood technology engineer to being a data scientist. Their impact on my life is immeasurable.

Lastly, I would like to say thank you to my close friends Pablo and Facundo who, besides giving me advise, lend me their company during countless evenings with asados and Fernet and Coke. To Dr. Robert Heigermoser who consistently supported me during my journey and did not let me forget that a PhD in the end is a marathon not a sprint.

Last but not least, I would love to express my sincere gratitude to my loving Family, my parents Roman and Eva, and my Brother Roman who gave me unlimited support and encouragement, and strengthened the importance to look after my health first. As their love for me would still exist with or without a PhD. Finally, to my wonderful girlfriend Ally who's love and care gave me a home so far away from home.

Abstract

AI is revolutionizing the technological landscape in medicine. A key application is the AI-driven summarization of clinical text, which facilitates the harmonization and curation of clinical data elements for common data models leading to improved understanding of population-level health. Population-level health is derived from aggregating patient-level information stored in unstructured electronic health records, often in the form of free-text clinical notes. As clinical text documents are information dense and written in highly complex clinical language, a model's ability to discern signal from noise becomes exceedingly more crucial. To enable models to identify relevant information in text documents, previous research has shown attention mechanisms and non-medical human-based rationales to be effective. Building on this foundation, this dissertation evaluated methods to optimize attention mechanisms and to effectively use human-based clinical rationales as additional supervision to improve the performance and interpretability of clinical text classification models. In particular, this dissertation shows the following: (i) the effective utilization of the reference information—initialization with external text code descriptions or encoding of code hierarchy of medical coding systems—contained by an attention mechanism's query matrix can improve model performance; (ii) extending an attention mechanism's receptive field with a flexible context window at the phrase-level leads to improved understanding of local linguistic information in clinical text; and (iii) utilizing human-based clinical rationales as additional supplementary training data can improve model performance and positively impacts model interpretability.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Significance

Artificial intelligence (AI) is transforming the technological landscape of medicine and oncology [120], providing an opportunity to create automated, self-learning systems that enhance the efficiency of clinical healthcare, medical research, and oncological patient care. In oncology, AI-driven clinical support systems have proven to enhance the accuracy of early detection methods, such as medical imaging analysis [141], assist in personalized medicine like genomic sequencing [114], and aid in rapidly summarizing vast amounts of clinical text. Rapid summarization of clinical text enables the timely identification of current cancer trends allowing for targeted research and population-level prevention strategies to foster public health [109, 41]. However, the current manual processes associated with collecting and assessing individual patient cases cause significant time delays [130], preventing real-time population-level cancer surveillance [70]. To address this issue, cancer surveillance programs like the Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries (NPCR) and the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) program are leveraging AI-based systems to automate the collection, preprocessing, information extraction, and harmonization of cancer data from electronic health records (EHRs) [21, 70, 109]. In 2016, the NCI partnered with the Department of Energy (DOE) and established the *Modeling Outcomes Using Surveillance Data and Scalable Artificial Intelligence for Cancer (MOSSAIC)* to develop AI-driven models based on natural language processing (NLP) and deep learning (DL) algorithms to alleviate the burden of certified cancer registrars from annotating large amounts of SEER-based electronic pathology reports. As part of these efforts, this dissertation project explores popular attention mechanisms in contemporary neural networks and clinical rationales to enhance the performance and explainability of AI systems, enabling the automatic retrieval of essential cancer information from unstructured electronic health records and ultimately improving cancer surveillance efficiency and patient outcomes.

Electronic health records (EHRs) are sets of structured and unstructured patient-level health data [139]. Structured data usually contains information on patient demographics, weight, height, laboratory tests, and medications [107]. Unstructured data, however, contains roughly 80% of the information stored in EHRs and comes as free-form text written by medical professionals [83]. Clinical documents such as electronic pathology reports or hospital discharge summary notes contain vital information on a patient's history, diagnosis, and treatment. As the generation of unstructured EHRs continues to rise, for example, the annual number of new cancer diagnoses and therefore cancer pathology reports approaches the two million mark [21, 130], the manual extraction of important data elements from clinical text becomes exceedingly challenging due to its time-consuming, costly, error-prone, and expertise demanding nature [135]. To overcome these barriers and automate the extraction of core patient-level disease information from clinical text, the biomedical informatics community have been heavily exploring natural language processing (NLP) and machine learning algorithms [47, 139].

Early state-of-the-art work utilized rule-based NLP [23, 79, 102, 144] and statistical ML algorithms such as boosting of weak learners, logistic regression, and support vector machines [101, 105, 124, 147] to retrieve information from clinical text. Clinical text, however, possesses inherent linguistic variability [122], leading to either unmanageable complexity in handwritten rules [100] or the need for manually engineered features that inadequately capture semantic relationships in clinical text [123]. To address these challenges, research shifted towards models deploying deeper architectures based on DL. The vast majority of these DL-based models utilize convolutional neural networks (CNNs) [13, 60, 62, 118], recurrent neural networks (RNNs) [86, 90, 93, 131, 155], or hybrid models that try to combine the strengths of both previously mentioned architectures [159, 165]. While these models are able to learn features directly from the raw text, with increasing depth of these architectures it becomes more difficult to identify the important parts of the input. To overcome this

issue, previous and current text classification models started to deploy a mechanism called attention.

Attention is a modular, unsupervised mechanism used to enable a model identify the important words in a text sequence [104]. Despite an initially slow adoption, attention became a pivotal structure in modern deep neural networks as it allowed for improved performance and explainability enabling state-of-the-art performances across many biomedical and clinical text classification tasks [48]. Interestingly, while previous and current state-of-the-art models deploy different approaches to encode the raw clinical documents, they virtually rely on the same types and implementations of attention mechanisms to accurately extract the salient features of an encoded clinical document. This lack of innovation in attention mechanisms suggests an overlooked opportunity to optimize and enhance the performance and explainability of modern text classification models.

Despite the significant improvements in text classification models through the implementation of attention mechanisms, recent work in non-clinical domains (e.g., movie reviews [158] or Wikipedia articles [33]) explored human-based rationales as a direct supervision signal on the important parts of a text sequence improving model performance and explainability [146]. Generally, rationales are concise, plausible explanations that adequately justify a document's ground-truth label. In the context of the extraction of critical cancer data elements, these rationales are annotated highlights in a given electronic pathology report which supports the associated ground-truth classification [26, 146]. While initial results with non-clinical rationales are promising, it remains unclear if human-based clinical rationales have the same impact on the clinical documents, and hence, on clinical text classification models.

Therefore, this dissertation research focuses on the development of novel attention mechanisms and the exploration of clinical rationales in order to improve the performance and explainability of clinical text classification models trained for the automated medical encoding of clinical documents. In particular, the focus lies on adapting adapting how attention mechanisms learn a suitable reference information

or how attention mechanisms can dynamically encode the input sequence local prior the comparison with the reference information. Importantly, the developed attention mechanisms must be (i) modular in design, (ii) independent of a model's text-encoder architecture (e.g., CNNs, RNNs, or Transformers), (iii) computationally efficient in terms of compute time and memory usage, and (iv) easy to implement into an existing codebase [119]. In contrast, the clinical rationales will be evaluated whether they should be used (i) as supplementary samples that enable models to discover new connections between input features and output labels; (ii) as fine-grained supervision, providing additional ground-truth labels at the word or sentence level; (iii) as guidance for internal model components, like attention mechanisms, to direct a model's focus towards rationale-like features. The primary task used to evaluate the developed attention mechanisms and clinical rationales is the information extraction of key cancer elements — site, subsite, laterality, histology, and behavior — from SEER-based electronic pathology reports. Finally, the dissertation will be centered around the following research questions:

1. How does the reference information in popular attention mechanisms impact model performance? Does the initialization strategy of the reference information matter? Can you encode the information on the code hierarchical of medical encoding systems via the reference information?

2. Can traditional word-level attention mechanisms be extended to accurately extract complex clinical language from clinical text by identifying longer, cohesive groups of words or phrases to improve the performance of clinical text classification models?

3. Can human-based clinical rationales improve the performance and explainability of attention-based clinical text classification models?

1.2 Clinical Text Datasets

1.2.1 Dataset 1: Electronic Cancer Pathology Reports

The National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) Program collects millions of cancer pathology reports to monitor and combat cancer among the American population [2]. Every cancer pathology report has a unique CTC; each CTC may be associated with one or more pathology reports. For each CTC, certified cancer registrars (CTRs) used all information at hand, e.g., electronic pathology reports, images, or laboratory results, to manually annotate the ground-truth labels of key data elements like primary cancer site, subsite, laterality, histology, and behavior. However, this annotation approach can lead to discrepancies between the ground-truth label and the content of individual reports, especially for CTCs with multiple reports. To minimize these mismatches, we excluded all CTCs annotated as *"Distant site(s)/node(s) involved,"* indicating metastasizing cancer according to the SEER summary stage classification scheme. Furthermore, CTRs additionally annotated a small subset of electronic pathology reports with clinical rationales that support the selected diagnosis for the cancer data elements site, laterality, histology, and behavior. The prediction of each of the five cancer disease data elements — site, subsite, laterality, histology, and behavior — is defined as a multiclass text classification task. Therefore, each model $f(w)$ with parameters w is tasked with the prediction of the most appropriate class $\hat{y}_{i,task} = softmax(f(w, x_i))$ from the task-specific label space $|L_{task}|$ of the set of N documents $D = \{(x_i, y_i)\}_{i=1}^{N}$. We followed the preprocessing procedure of the raw cancer pathology reports described in Gao et al. [49] to create a cleaned, deidentified, and tokenized dataset. The specific datasets used for the respective set of experiments are described in each separate chapter.

1.2.2 Dataset 2: Hospital Discharge Summary Notes

Medical Information Mart for Intensive Care III (MIMIC-III) is a publicly available clinical database that contains deidentified EHRs (e.g., clinical notes, vital signs, laboratory measurements) for more than 40,000 patients [69]. Each unique patient ID is associated with at least one hospital admission ID. Every unique hospital admission ID was annotated by trained human encoders with a subset of the ICD-9 diagnosis and procedure codes. This subset of labels $y_i \in \{0,1\}^l$ associated with document i from the label set $|L|$ of the set of N documents $D = \{(x_i, y_i)\}_{i=1}^{N}$ has to be predicted by only using the hospital discharge summary notes and available addenda related to a respective hospital admission ID. This prediction problem is defined as a multi-label document classification task. The procedure from Mullenbach at el. [99] was adopted for the preprocessing of the clinical text, which tokenized each sequence, lower-cased all tokens, and removed non-alphabetic tokens or tokens that occurred less than tree times in the corpus.

1.3 Interplay between Data Availability and Architectural Improvements

The collaboration between the NCI, the participating SEER central cancer registries, and the DOE enables access to an unprecedented amount of electronic pathology reports containing crucial patient-level cancer information. The appropriate retrieval of patient-level cancer information can enhance population-level cancer surveillance and improve patient outcomes. This, in turn, facilitates more timely and targeted cancer research and prevention strategies [109]. By developing these strategies for specific demographic groups—defined by factors such as sex, age, or geographical location—and linking them to specific cancer types, including rare cancers, patient outcomes can be significantly improved [41]. This information extraction depends on two primary factors: (i) the availability of efficient and accurate AI-driven systems to

automate the retrieval of vital information on key cancer data elements and (ii) the availability of a rich data source that sufficiently represents the disease occurrence among the American population.

DL-based algorithms require large amounts of samples across all classes to appropriately map the input to the correct class [8]. However, due to inherent variabilities in the occurrence of different cancer types among the American population, clinical text datasets such as the SEER dataset used in this work 1.2.1 will often show significant imbalance in their class distributions. In particular, rare cancer types (see Table **??**) may only have a small number of examples in the dataset exacerbating the process of a model to accurately extract a representative set of features that describe each class in a generalized way without memorizing individual samples. To overcome this issue, programs like SEER need to specifically look for cancer patients diagnosed with rare cancers and make them accessible to computational cancer research. Besides increasing the number of available samples across all classes, it is important to not neglect the continuous development of AI-based systems designed for the automated extraction of key cancer data elements from electronic pathology reports. Specifically, these models must be able to appropriately classify majority classes at the cost of appropriately predicting minority classes.

Nevertheless, a successful and timely execution of population-level cancer surveillance necessitates a positive interplay of both factors. Efficient state-of-the-art AI-based algorithms are limited in their usefulness without access to a rich data source that contains sufficient samples for all types of cancer, but especially for rare cancer types. However, the availability of a rich data source alone is not sufficient for cancer surveillance without the availability of efficient, accurate, and explainable AI-based algorithms to retrieve the cancer information hidden in millions of electronic pathology reports. Finally, only simultaneous efforts at both fronts, the collection of reports and the accurate extraction of cancer information, will lead to enhancements, while any small improvements in this space will have significant impact on patient-level outcomes.

1.4 Literature Review

1.4.1 Natural Language Processing

Natural language processing (NLP) is the intersection of computer science and linguistics concerned with teaching computers how to understand human languages [73]. In the 1950s, the first primal form of NLP-related research was done in the field today known as machine translation [100]. Despite the consensual agreement that the translation from one language to another is important, early machine translation models couldn't perform as expected leading to a dramatic halt in this type of research [63]. Additionally, a major barrier in performance for earlier NLP algorithms was their inability to understand language syntax, i.e., models could not differentiate between grammatically correct and incorrect sentences. Noam Chomsky, widely respected as the *father of modern linguistics* [34], introduced the theory of universal grammar which postulates that all languages with some degree of variation follow a formal set of structures and rules [32]. With the development of Chomskyan Linguistics, early NLP has seen tremendous research activities and success across a wide range of domains and task. At this time, NLP research heavily relied on handwritten rules via regular expression indicating good performance on specific tasks; however, improving such models required the addition of a more complex set of rules which becomes increasingly more difficult.

To overcome the necessity of using handwritten-rules and enabled by the continual advancement in computing technologies [110], research gradually moved towards a class of algorithms known as machine learning (ML). Machine learning algorithms such as Naïve Bayes Classifier, Hidden Markov Models, or Support Vector Machines utilize statistical inference to automatically learn rules through the analysis of abundant real-world data [100]. The quality of these ML-based rules can be improved by adding more data to generate a richer target distribution leading to better model performance. Unfortunately, a big drawback of these algorithms is that they require

input that was created through manual feature engineering. Early feature engineering techniques such as bag-of-words [55] or TF-IDF [132] aimed at creating a single vector representation to describe a document. Both techniques create large and sparse vectors where each element represents the information for a given word and document ultimately resulting in a loss of word order and semantics and large dimensionalities [30]. Despite the success of traditional ML algorithms, the development of novel model architectures (e.g., neural networks, attention mechanisms, and transformers) and newer, faster accelerators (i.e., GPUs) caused their displacement by the emergence of a new class of algorithms known as representation learning or deep learning (DL).

Deep learning is a subfield of ML and comprises a class of algorithms that can directly capture salient features from raw, unprocessed data avoiding prior manual feature engineering of the documents [81]. To achieve this, modern neural networks represent each word in a document using an embedding layer which internally maps a specific word to its corresponding word embedding [18]. Word embeddings are complex, dense vector representations with a fixed dimension that are either initialized randomly or with pretrained weights. These word embedding layers are often initialized with vector representations pretrained on the text using algorithms like Word2Vec [96] or GloVe [108] which has shown to improve performance and convergence compared with random initializations [75]. Once all words of a document are transformed to their respective word embedding vector, neural networks encode the meaning of a word relative to either the local or entire sequence using specific text-encoder architectures. Typical text-encoder architectures in deep learning models are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers.

1.4.2 Fundamental Deep Learning based Text-Encoder Architectures

CNNs utilize one or multiple convolution layers, each equipped with a sliding window that moves across the text, to learn short-range sequences of n consecutive words or n-grams [74]. These n-grams contain local information focusing on specific keywords relevant to the task at hand; however, this architecture fails at capturing contextual linguistic information which can span across the entire document. In contrast, RNNs with its two most popular variants, the long short-term memory (LSTM) [57] and the gated recurrent unit (GRU) [31] neural network, use their gating mechanisms to model long-term dependencies for sequential data without suffering from the exploding or vanishing gradient problem [43]. Both variants find frequent use in NLP-tasks since there is no definite consensus in the literature on the superiority of one architecture over the other [154].

In 2017, Vaswani et al. [142] introduced the most recent iteration of a fundamental DL text-encoder architecture to NLP called the transformer; transformers are the fundamental backbone of modern large language models like ChatGPT, Llama, or Claude. The transformer solely consist of multiple layers of self-attention. Self-attention aims at finding the important parts in a text sequence by comparing the input with itself resulting in a quadratic complexity. Due to this quadratic complexity, early transformer-based models like BERT had to limit the processable maximum document length to only 512 tokens to avoid running into memory issues [36]. Since clinical documents tend to be longer in nature, BERT-based research either had to trim documents neglecting remaining information in the document entirely or to split the documents into equally long chunks. Each chunk would then be encoded independently by BERT causing the model to not learn long-range dependencies ranging across chunks. To overcome these issues with self-attention, research focused on reducing the memory consumption of attention mechanisms in transformers to overcome the sequence length limitation [17]. One technique proposed was to

implement sparse self-attention allowing the processing of long clinical documents up to 4096 tokens [84]. Due to the success of the transformers on NLP tasks with clinical text, the popularity of attention mechanisms increased resulting in becoming a pivotal structure in modern NLP-based models.

1.4.3 Attention Mechanisms

Attention mechanisms mimic a human's ability to focus, a cognitive process that selectively perceives the incoming information stream [48]. Translated to computer science, attention is an unsupervised mechanism that dynamically generates soft weights, so-called attention scores, to indicate the relative importance of a word relative to all remaining words in a text document [22]. From its inception in the NLP in 2014 [11] to being the sole component of the transformer architecture, modern attention mechanisms [142] have shown to tremendously improve model performance a wide range of NLP tasks like aspect-level sentiment analysis, opinion mining, or text classification. Furthermore, the internally generated attention scores have shown to be useful for the interpretation, often synonymously used for explanation, of a model's predictions. In particular, many studies proposing attention-based models to classify clinical text present examples showing extracted phrases with words that are highlighted in color corresponding to their respective attention score. Alternatively, attention scores can be used to rank the importance of individual or sequences of words allowing the extraction of the top k text snippets supporting a given predictions.

Modern attention mechanisms follow the implementation of Vaswani et al. [142] that utilizes three primary inputs — the key-value pair matrices and the query matrix — to identify the important parts of an input text sequence. In text classification, the keys (K) and values (V) are two sequences encoding the salient features of an input text sequence, while the query matrix (Q) encodes a learned task-specific reference information, e.g., medical concept on a disease or specific cancer type. The mechanism consists of two operations: (i) a matrix multiplication between K and Q resulting in

energy scores which indicate the strength of the global relationship between a word and the medical concept and (ii) a softmax operation which normalizes the energy score to create attention scores. These attention scores indicate the importance of each word relative to all remaining words in the document.

Although there is a rich architectural landscape for different attention mechanisms, contemporary state-of-the-art clinical text classification models converged to mainly using the following three types of attention mechanisms in their architectures: (i) self-attention used to encode the contextual meaning of each word by comparing the sequence with itself [36, 50]; (ii) target-attention which compares each encoded word with a task-specific reference information [50]; and (iii) label-attention which extends the idea of target-attention and compares each encoded word to reference information specific to each class associated with the task [38, 44, 88, 87]. Importantly, self-attention mechanisms are usually utilized to encode a text sequence using contextual information, whereas, target- and label-attention are primarily used to compare the final encoding of a document, sometimes even generated with the self-attention mechanism, against task-specific reference information.

1.4.4 Attention-based Models for the Automated Classification of Clinical Text

The information extraction from clinical text is a challenging task [4]. Clinical text differs from conventional text as it is written in telegraphic, ambiguous clinical language filled with jargon, technical terms, and abbreviations [139] and often spanning thousands of words. To overcome these challenges, the vast majority of contemporary models designed to classify clinical text rely on deep neural networks that incorporate attention mechanisms. This attention-based deep neural networks have shown significant performance improvements over the past years and represent the current state-of-the-art in two primary clinical text classification tasks in the current literature, the information extraction of key cancer data elements from

electronic pathology reports [50, 148] and the automated encoding of hospital discharge summary notes with medical codes [129, 99, 143].

Due to regulatory restrictions concerning patient privacy, there are only a limited number of studies proposing models incorporating attention on the information extraction from pathology reports. The first notable work in this space explored hierarchical attention networks (HAN) that deploys RNNs to encode a document and uses an early attention layer to identify important words [52]. After several iterations, Gao et al. optimized their initially proposed (HAN) by substituting the RNN-based text-encoder with two parallel self-attention mechanisms [51], and later consolidated the self-attention mechanisms into a single layer [50]. Wu et al. [148] investigated an attention-based graph convolutional neural network which showed promising performance on the information extraction from cancer pathology reports and outperformed another notable model frequently used in this space, the multi-task convolutional neural which does not use attention [6].

For the latter task, researchers have access to a publicly available hospital discharge summary notes which explains a significant larger body of work. Since in this text classification task, multiple labels are assigned to a document, the majority of proposed models incorporate the label-attention mechanism as it can learn label-specific reference information. However, these models primarily differ in their selected text-encoder architecture. Earlier work utilized either CNNs [99, 24, 67, 82, 60, 88] or RNNs [129, 143, 38], whereas, more recent work started to utilize transformer-based text-encoder architectures [44, 56, 20, 61, 87].

1.4.5 Human-Based Rationales

Modern clinical text classification models heavily rely on unsupervised attention mechanisms to identify important text passages relevant to assigned medical diagnoses in clinical documents. Despite attention's strong capabilities in finding these connections, its unsupervised approach can sometimes identify words or sequences

as important that seem irrelevant to the associated diagnosis. To address this mismatch, researchers in non-medical text classification have explored using human-based rationales as additional supervision [158, 26, 116].

To better understand the role of rationales in addressing this issue, it's important to define what rationales are and how they're used. Rationales are concise, plausible explanations that adequately justify a document's ground-truth label, such as a clinical diagnosis [26, 146]. They can be either abstractive (free-form text created by generative models) or extractive (text highlights retrieved by a model or human) [158, 161, 28, 157]. Researchers have leveraged rationales as a form of prior domain knowledge to improve model performance and interpretability [136, 9, 162]. While recent work has explored large language models for generating abstractive rationales [78, 121], this dissertation focuses on utilizing extractive rationales to enhance discriminative clinical text classification models.

Utilizing human-based rationales aims to enhance a model's ability to identify key text elements, thereby improving both performance and explainability [146]. Explainability can be categorized as either faithful, accurately representing a model's internal decision-making process, or plausible, designed to be easily understood by humans [136]. While human-based rationales are often used as plausible explanations, attention mechanisms are typically used to provide faithful explanations. However, there has been considerable debate regarding the suitability of attention for providing faithful explanations for a model's predictions [37]. This debate centers on whether attention mechanisms, despite their use for faithful explanations, can truly provide accurate insights into a model's decision-making process, particularly in complex tasks like clinical text classification.

1.5 Research Objectives

This dissertation will focus on developing novel attention mechanisms and the exploration of clinical rationales to improve the performance and explainability by

allowing clinical text classification models to better focus on the important parts of a clinical document. This dissertation's novel contributions are as following:

1. A comprehensive evaluation of the initialization strategies of the reference information used by popular attention mechanisms (i.e, target- and label-attention) in current clinical text classification models filling a missing piece in the literature on how to effectively utilize attention mechanisms.

2. The development of a novel label-attention mechanism that can learn and incorporate the inherent hierarchical structure of medical encoding systems (e.g., ICD-9) via the enrichment of the implicitly learned label-specific reference information without requiring an additional auxiliary network. This hierarchical attention-mechanism first learns a mapping between the salient features of the high-level categories from the clinical text to then inform the label-specific reference information associated with the low-level codes.

3. The development of a novel attention mechanism that can learn salient word-level and phrase-level context features in parallel from clinical text improving performance of a text classification model. This phrase-level attention mechanism was designed as a distinct module to extend any type of text-encoder architecture (i.e., CNN, RNN, or Transformers).

4. The investigation of human-based clinical rationales as an additional supervision signal to improve transformer-based clinical text classification models performing the automated medical encoding of critical cancer data elements. These clinical rationales represent plausible explanations for the selected cancer data element and should help the model to better identify rationale-like features improving model performance and interpretability.

1.6 Dissertation Outline and Timeline

This research project started in February 2022 and was completed in November 2024. The primary expectation of this research project was the development of three manuscripts on improving the automated medical encoding of clinical documents, e.g., the extraction of important cancer data elements from electronic pathology reports. The dissertation utilizes these three manuscripts as its three core chapters (2, 3, 4). Chapter 5 provides a conclusion to this dissertation.

Chapter 2 contains the manuscript titled *Attention Mechanisms in Clinical Text Classification: A Comparative Evaluation* that was created from February 2022 to May 2023 and published in the *IEEE Journal for Biomedical and Health Informatics* in January 2024. This work focused on understanding the nuances between the initialization of an attention mechanism's reference information—random or pretrained—and incorporation of the code hierarchy information of medical coding systems via attention mechanisms.

Chapter 3 presents the work titled *Deformable Phrase Level Attention: A Flexible Approach for Improving Ai Based Medical Coding* that was created from June 2023 to March 2024. The manuscript presented in Chapter 3 was submitted to the *Journal of Artificial Intelligence in Medicine* and is currently under review. This work proposes a novel attention mechanism that enables a model to dynamically identify important information at the phrase-level with flexibly sized context windows.

Chapter 4 contains the work titled *Can human clinical rationales improve the performance and interpretability of clinical text classification models?* which was performed from April 2024 to October 2024. This work is intended to be submitted to the *Journal of the American Medical Informatics Association* and is currently under review. The manuscript contains work on how clinical rationales can improve clinical text classification models performing the automated medical encoding of electronic pathology reports with the primary cancer site.

Chapter 2

Attention Mechanisms in Clinical Text Classification: A Comparative Evaluation

2.1 Disclosure Statement

A version of this chapter was originally published in IEEE Journal of Biomedical and Health Informatics:

C. S. Metzner, S. Gao, D. Herrmannova, E. Lima-Walton and H. A. Hanson, "Attention Mechanisms in Clinical Text Classification: A Comparative Evaluation," in IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 4, pp. 2247-2258, April 2024, doi: 10.1109/JBHI.2024.3355951.

Author's contributions are as follows: CM designed and executed the study; SG, HH, and DH contributed to the experimental design and analysis of the results. ELM performed qualitative evaluation. CM organized, programmed, and performed all experiments. CM wrote the manuscript. All authors read, reviewed, and approved the final manuscript.

No revisions to this chapter have been made since the original publication.

2.1.1 Abstract

Attention mechanisms are now a mainstay architecture in neural networks and improve the performance of biomedical text classification tasks. In particular, models that perform automated medical encoding of clinical documents make extensive use of the label-wise attention mechanism. A label-wise attention mechanism increases a model's discriminatory ability by using label-specific reference information. This information can either be implicitly learned during training or explicitly provided through embedded textual code descriptions or information on the code hierarchy; however, contemporary studies arbitrarily select the type of label-specific reference information. To address this shortcoming, we evaluated label-wise attention initialized with either implicit or explicit label-specific reference information against two common baseline methods—target-attention and text-encoder architecture-specific methods—to generate document embeddings across four text-encoder architectures—a convolutional neural network, two recurrent neural

networks, and a transformer. We also present an extension of label-wise attention that can embed the information on the code hierarchy. We performed our experiments on the MIMIC III dataset, which is a standard dataset in the clinical text classification domain. Our experiments showed that using pretrained reference information and the hierarchical design helped improve classification performance. These performance improvements had less impact on larger datasets and label spaces across all text-encoder architectures. In our analysis, we used an attention mechanism's energy scores to explain the perceived differences in performance and interpretability between the text-encoder architectures and types of label-attention.

2.2 Introduction

Roughly 80% of the information stored in EHRs is unstructured data and comes as free-form text written by healthcare professionals [83]. However, the manual retrieval and classification of this information are infeasible because these activities are time-consuming, cost-intensive, error-prone, and require significant domain knowledge [135]. To overcome these barriers and automate the classification of clinical documents, the biomedical informatics community leverages machine learning, natural language processing (NLP), and deep learning-based methods. Particularly, contemporary NLP models use the popular domain-agnostic building block located at specific positions in the architecture of deep neural networks called attention.

Attention mechanisms dynamically manage the perceived information stream by using soft weights to indicate the relevance of a token to a task within a text input sequence[12]. The introduction of the Transformer [142] further increased the popularity of attention mechanisms because it enabled state-of-the-art performance across multiple NLP tasks while relying solely on multiple self-attention layers. As a result, current state-of-the-art clinical document classification models incorporate different types of attention mechanisms. For example, self- and target-attention have been deployed by the HiSAN model to classify cancer pathology reports [50],

whereas models that perform the automated encoding of hospital discharge summaries primarily utilize label-wise attention [99, 143, 87].

Label-wise attention's popularity in extreme multi-label scenarios stems from its ability to incorporate label-specific auxiliary information [24, 27]. Auxiliary information can be either implicitly learned through randomly initialized embeddings or explicitly provided by using embeddings of textual code descriptions or information on the code hierarchy [68]. Selecting the appropriate type of auxiliary information for the task and dataset at hand is nontrivial; however, in most studies, this selection is often arbitrary. Furthermore, given the often-existing linguistic mismatch between the formally defined code descriptions and the informally written clinical documents [129], it is important to quantify the differences between the outputs of different types of text-encoder architectures with differently initialized label-attention mechanisms.

In this work, we evaluate label-attention mechanisms that incorporate either implicit or explicit auxiliary information across multiple text-encoder architectures — convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers — in the multi-label classification setting using hospital discharge summary notes. We also compare the performance of our label-attention mechanisms against target-attention and other common methods for generating document embeddings. Our contributions are as follows:

- We perform the first comprehensive comparative study on the label-attention mechanism in clinical text classification with a focus on improving the automated encoding of clinical documents with medical codes (i.e., ICD-9).
- We examine the effect of different attention mechanisms on CNN-, RNN-, and transformer-based text-encoder architecture-specific document embeddings.
- We show that initializing an attention mechanism's reference information with explicit label-specific auxiliary information improves classification performance.

- We are the first to demonstrate a performance improvement in clinical document classification by incorporating information on code hierarchy solely via the attention mechanism with minimal increase in compute time and without requiring an additional neural network.

2.3 Related Work

Attention mechanisms are universally used in contemporary deep learning models and significantly improve performance in a wide range of NLP tasks, including aspect-level sentiment analysis [71], opinion mining [113], neural machine translation [45], and text classification [152]. Despite the ubiquity of attention mechanisms in deep learning, only a few studies in the NLP domain have empirically compared their effects on the performance of deep neural networks. Kardakis et al. [71] investigated global-, self-, and hierarchical-attention mechanisms in RNNs by performing sentiment analysis of movie reviews that showed an average increase in accuracy through attention by two points. Jain et al. [65] and Feucht et al. [44] analyzed a single type of attention across multiple encoder architectures to predict ICU-related tasks (e.g., readmission, medical encoding of notes) by using MIMIC-III clinical notes. Jain et al. [65] tested target-attention, which utilizes a single query that contains the reference information about the entire label space, to find important parts within documents and reported accuracy improvements in CNNs and Bi-LSTMs. Feucht et al. [44] investigated how the performance of different transformers (e.g., BERT, Hierarchical-BERT, Longformer) is impacted by similar label-attention mechanisms with pretrained embeddings of textual code descriptions. Although their work is similar to ours, the underlying relationships between the different types of text-encoder architectures and attention mechanisms remain unexplored.

A noticeable trend in deep learning models used to perform automated medical encoding of clinical documents is the reliance on an attention mechanism that can incorporate label-specific information. Early work by Shi et al. [129] highlighted

the benefits of incorporating textual code descriptions in their attention mechanism to assign ICD-9 codes to sentence-level diagnosis descriptions. Their attention mechanism creates a single weighted document representation that contains the similarities between a diagnosis description and all considered ICD-9 code description embeddings. Mullenbach et al. [99] argued that using a single document representation is insufficient to assign multiple medical codes to a clinical note and proposed a label-attention mechanism instead. Label-attention generates one document representation for each label in the label space to capture locations in the text that are semantically relevant to the specific label by using queries with label-specific reference information. Contemporary state-of-the-art models utilize label-attention as a critical component in their architecture but differ in their approach to initializing the query matrix randomly or pretrained. Randomly initialized queries are learned during training [99, 143, 88], whereas pretrained queries use label description embeddings as a starting point and are fine-tuned during training [99, 24]. For example, Liu et al. [87] proposed the hierarchical label-wise attention transformer model that uses randomly initialized label-attention on token- and chunk-level data to extract the most informative features from a clinical document, thereby achieving state-of-the-art performance on MIMIC-III-50. Despite label-attention's widespread use, the different architectures of the proposed models prevent a fair evaluation of both query initialization strategies. Therefore, we perform the first comparative evaluation of label-attention by following the implementation of Mullenbach et al. [99].

Medical encoding systems have an intrinsic hierarchical structure that provides information about the relationships between high-level categories and low-level codes. A large body of work on the subject contains diverse methods to exploit these hierarchical relationships to improve model performance [68]. Some studies use label co-occurrence or label correlation during the training process to initialize specific parts of the neural network by introducing bias and exploiting the parent-child relationships between the labels [77, 14]. Others use graph neural networks (GNN) to learn the hierarchical relationships of the label space [29, 164]. Xu et al. [151] combined label

correlation and a GNN to learn a precise representation of the hierarchy. Shen et al. [127] instead used weak supervision to create a hierarchical structure of the label space. In particular, they approached the problem from the perspective of a layman trying to classify an unknown set of documents. They argue that laymen would first classify the documents with high-level categories before selecting more specific low-level labels. This way, knowledge about high-level categories will enrich the available information for predicting the low-level labels associated with a document. We adopted their idea of a layman approach to this problem in our implementation of the hierarchical label-attention mechanism and show that it is possible to incorporate the hierarchical structure in a simple attention mechanism.

2.4 Materials and Methods

2.4.1 General Document Classification Attention Model

Many deep learning models incorporate different attention mechanisms within different parts of their architectures, thereby making it difficult to perform a direct and fair comparison of the impact of attention mechanisms on a model's performance [48]. With this fair comparison in mind, we designed a general attention-based document classification model (Fig. 2.1) that consists of the following layers: (i) an embedding layer; (ii) a text-encoder architecture; (iii) an attention layer; and (iv) a linear output layer. We describe each layer in more detail in the following sections.

2.4.2 Embedding Layer

The first layer of each model is a pretrained word-embedding layer that takes a text sequence as input, $X = [x_1, x_2, \ldots, x_N]$, where N is the number of words. The embedding layer transforms each word to its corresponding dense vector representation with dimension d_e, which results in a word-embedding matrix, $X_E \in \mathbb{R}^{N \times d_e}$. For the CNN- and RNN-based models, we initialize the word-embedding layer

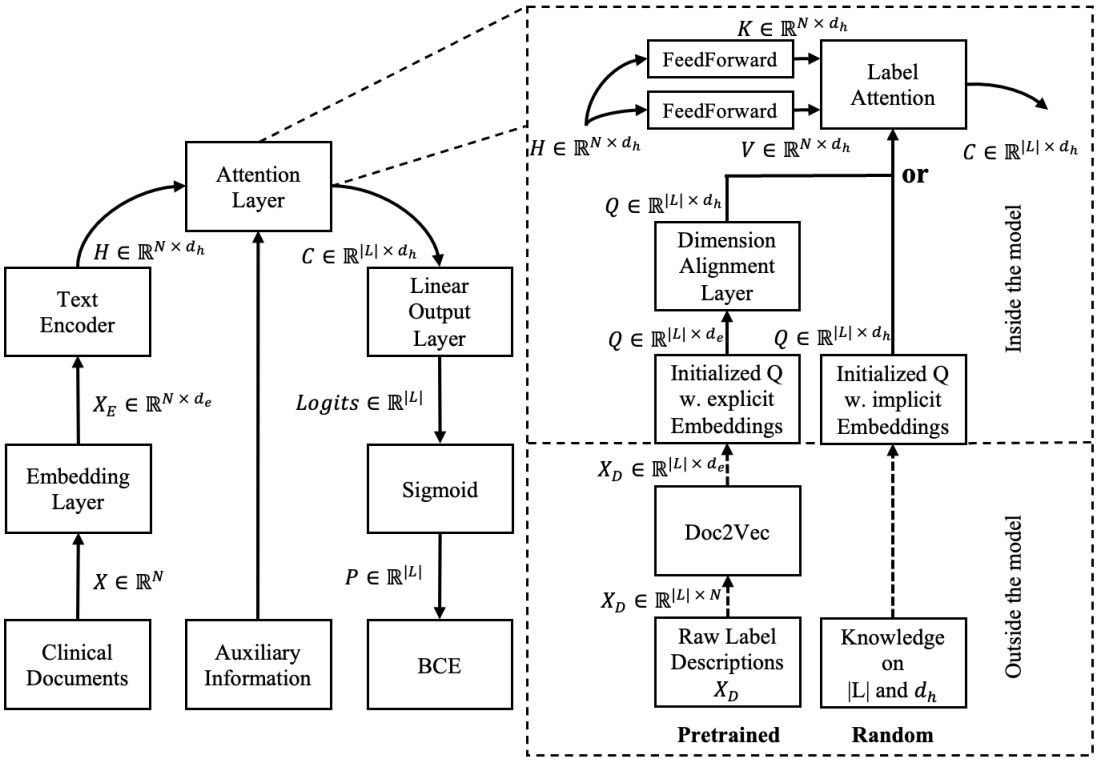


Figure 2.1: Structure of the general document classification attention model. The text-encoder segment represents a CNN-, RNN-, or transformer-based architecture. The label-attention layer utilizes implicit (i.e., random embeddings) or explicit (e.g., embedded textual code descriptions) auxiliary information. The multi-label classification models use a sigmoid activation function after the output layer and the binary-cross-entropy objective function.

with pretrained embeddings of each word in the vocabulary generated with Word2Vec [96]; the vocabularies and word embeddings were created from the training documents of each dataset. In contrast, the pretrained transformer model comes with a byte-pair-encoding vocabulary based on the corpus associated with the model during pretraining [17].

2.4.3 Base Text-Encoder Architectures

We consider four base text-encoder architectures in this study: (i) CNN, (ii) bi-directional long short-term memory network (Bi-LSTM), (iii) gated recurrent unit network (Bi-GRU), and (iv) transformer-based Clinical-Longformer (CLF). We limited the scope of our study to these four text-encoder architectures as they represent the most-common and fundamental architectures in the contemporary clinical text classification literature [89]; we selected the CLF as the representative of the transformer class as it was pretrained on clinical text and can process up to 4096 tokens, which is important for longer clinical documents.

Generally, each encoder type takes the word-embedding matrix, X_E, as the input and learns a latent document representation with dimension d_h, which results in a latent document matrix, H. A dropout layer with probability p is connected in series to prevent overfitting [133]. The model passes the dropout layer's output to an attention mechanism or other common method used to generate document embeddings.

Convolutional Neural Network (CNN)

The first encoder architecture is a shallow CNN that follows the implementation by Y. Kim [74]. The CNN passes the learned word-embedding matrix X_E to three parallel 1D convolution layers with d_h filters and window sizes of three, four, and five tokens, thereby learning multiple n-grams that contain relevant task-specific information.

Following the ReLU activation, the three outputs are padded and concatenated, which results in the latent document matrix $H \in \mathbb{R}^{N \times 3d_h}$.

Recurrent Neural Network (RNN)

The next text-encoder architectures are two popular RNNs: the Bi-LSTM [57] and the less complex Bi-GRU [31]. Each bi-directional, two-layered RNN model takes the word-embedding matrix X_E as the input and learns a latent document matrix $H \in \mathbb{R}^{N \times 2d_h}$ with $2d_h$ representing the bi-directional pass over the size of the hidden states.

Transformer-based Clinical Longformer (CLF)

The final text-encoder architecture is the transformer-based CLF [84]. The CLF takes the word-embedding matrix X_E as the input and learns a latent document representation $H \in \mathbb{R}^{N \times d_h}$ with hidden dimension $d_h = 768$.

2.4.4 Attention Mechanism

In text classification, attention mechanisms guide a neural network's prediction process by assigning attention weights to each token in the input text sequence [104]. Modern attention mechanisms usually follow the design proposed by Vaswani et al. [142], which utilizes three inputs—the key-value pair matrices and the query matrix. The keys, K, and values, V, are two learned sequences that encode the key features of the latent document representation, H, whereas the queries, Q, are a reference to the information that the model is searching for within the input [22]. The mechanism uses K and Q to find relationships between the input text sequence and the reference information. We investigated if the type of initialization of Q, either random or with pretrained embeddings, influences model performance. Specifically, we evaluated the popular label-attention mechanism [99] in the single- and multi-head

setting and proposed a new hierarchical adaption of the label-attention mechanism that can incorporate the hierarchical relationships of the label space.

Label-Attention

Label-attention enables the model to learn a single document embedding for each label in the label space, $|L|$, thereby increasing a model's discriminatory ability [99]. In our implementation, we pass the latent document representation H into two separate position-wise feed-forward layers with the weights $W_K \in \mathbb{R}^{d_h \times d_h}$ and $W_V \in \mathbb{R}^{d_h \times d_h}$, biases b_K and b_V, kernel size and stride of one, and an Exponential Linear Unit (ELU) activation function to generate the key-value pair matrices $K \in \mathbb{R}^{N \times d_h}$ (2.1) and $V \in \mathbb{R}^{N \times d_h}$ (2.2). The query matrix $Q \in \mathbb{R}^{|L| \times d_h}$ is a trainable embedding matrix initialized either randomly (random label-attention) or with pretrained embeddings of textual code descriptions (pretrained label-attention). Textual code descriptions vary in length; however Q consists of fixed length embeddings. The Doc2Vec algorithm [80] is a natural solution to this issue, as it generates fixed-length embeddings with size d_e; each description represents its own document tagged with its respective medical code as the label. Each element of Q is a query and represents the reference information for a specific label. Label-attention uses the three inputs, K, V, and Q, to compute the label-specific context vector (2.3). We used a 1D convolution layer as a dimension alignment layer to map the resulting pretrained query embeddings with size d_e to d_h to avoid dimension mismatch between K and Q. The task-specific textual code descriptions were retrieved from ICD-9 [145].

$$K = ELU(FeedForward(H, W_K) + b_K) \tag{2.1}$$

$$V = ELU(FeedForward(H, W_V) + b_V) \tag{2.2}$$

$$C(Q, K, V) = softmax(\frac{Q \cdot K^\top}{\sqrt{d_h}})V \tag{2.3}$$

The scaled matrix product (i.e., compatibility function) uses K and Q to compute the energy scores, $E \in \mathbb{R}^{|L| \times N}$, which indicate the strength and direction of the association between the hidden representation of a token and the label-specific reference information in Q. The softmax operation (i.e., distribution function) retrieves the label-specific attention scores, $A \in \mathbb{R}^{|L| \times N}$, and the attention scores are subsequently applied to V to generate the label context matrix, $C \in \mathbb{R}^{|L| \times d_h}$, for a given document.

Hierarchical Label-Attention

This study proposes an extension of label-attention (Fig. 2.2) that can incorporate the code hierarchy of medical encoding systems. Hierarchical label-attention allows us to exploit the interdependence between the less specific but easier-to-predict high-level and low-level classifications. We argue that performing label-attention on two hierarchy levels enriches the downstream information and will lead to improved model performance. The hierarchical label-attention first generates context vectors, $C_{Cat} \in \mathbb{R}^{|L_{Cat}| \times d_h}$, for all high-level categories, $|L_{Cat}|$, associated with the label space by using label-attention described in Section 2.4.4 and the three input matrices, K, V, and, $Q_{Cat} \in \mathbb{R}^{|L_{Cat}| \times d_h}$. Q_{Cat} contains random or pretrained reference information for the high-level categories. Subsequently, we enrich the low-level label query matrix $Q \in \mathbb{R}^{|L| \times d_h}$ with the high-level category information in C_{Cat} (2.4). Specifically, we add the context vector $c_{cat,c}$ of the high-level parent category c to each related low-level label query $q_{l,c} \in L_c$, where L_c refers to all labels associated with category c, to obtain an enriched label query q_l for label l for all task-specific categories L_{cat}:

$$q_l(l_c, c) = q_{l,c} + c_{cat,c}$$
$$\forall 1 <= q_{l,c} <= L_c \quad \text{and} \quad \forall 1 <= c <= L_{cat}. \tag{2.4}$$

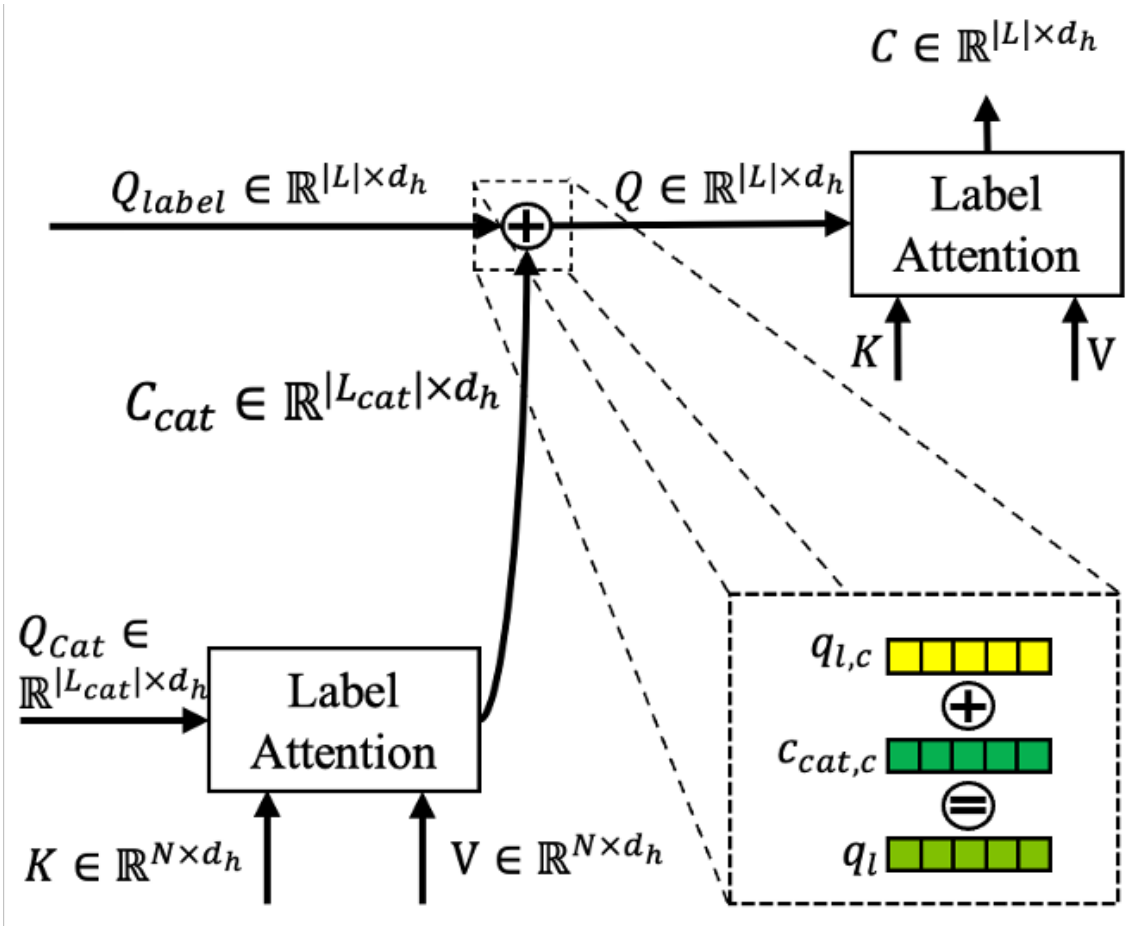


Figure 2.2: Flow of our hierarchical label-attention mechanism. Both K and V are generated from the encoder output by using a feed-forward layer with ELU activation. Both Q and Q_{cat} are initialized with random or pretrained embeddings. The label query vector $q_{l,c}$ is enriched with information on the code hierarchy by adding the parent category context vector $c_{cat,c}$.

We retrieved the high-level categories directly from the medical encoding system of each classification task. For the ICD-9 codes, we used the high-level chapters as the categories (e.g., *Infectious and Parasitic Diseases*) [145]. Our hierarchical label-attention design was inspired by a proposed attention-via-attention mechanism [48].

Multi-Head Label-Attention

We implemented a multi-head variant of the label-attention and hierarchical label-attention mechanisms. Multi-head attention [142] uses h attention heads to linearly project h different parts of the Q, K, and V matrices in parallel (2.5). The output is then concatenated and projected by using $W_O \in \mathbb{R}^{(h \times d_v) \times d_h}$.

$$\begin{aligned} Multihead(Q, K, V) &= Concat(head_1, ..., head_h)W_O \\ \text{where} \quad head_i &= Attention(QW_{i,Q}, KW_{i,K}, VW_{i,V}) \end{aligned} \tag{2.5}$$

The query, key, and value projection matrices have the shape of $W_{i,Q} \in \mathbb{R}^{d_h \times d_k}$, $W_{i,K} \in \mathbb{R}^{d_h \times d_k}$, $W_{i,V} \in \mathbb{R}^{d_h \times d_v}$, and $d_k = d_v = d_h/h$.

Baseline Methods

We tested the validity of the label-attention mechanism against two baseline methods: an architecture-specific baseline method for generating document embeddings and an alternative-attention mechanism called *target-attention*. For the first baseline, we substituted label-attention with a max-pooling operation over the logits after the output layer for CNN-based models. We used the average hidden state space over all time steps as the input to the output layer for both RNNs. Lastly, the transformer-typical baseline method utilizes the hidden representation of the first token of the encoded sequence as the input to the classification layer. Target-attention deploys a

query matrix, $Q_T \in \mathbb{R}^{1\times d_h}$, with a single element that attempts to capture the task-specific reference information on the label space and generates only a single context vector for prediction, $C_T \in \mathbb{R}^{1\times d_h}$ (2.6) [50].

$$C_T(Q_T, K, V) = softmax(\frac{Q_T K^\top}{\sqrt{d_h}})V \tag{2.6}$$

2.4.5 Output Layer

Given the generated context vector c_l, we compute the prediction probability of label l by using projection weights w_l and bias b_l of a linear layer, with $W \in \mathbb{R}^{|L|\times d_h}$ and $B \in \mathbb{R}^{|L|\times 1}$ (2.7), followed by the sigmoid activation function.

$$\hat{y}_l = act(w_l^\top c_l + b_l) \tag{2.7}$$

2.4.6 Training

We trained the model parameters by using the Adam optimizer and minimizing the the binary-cross-entropy loss for the multi-label experiments. We used a linear learning rate scheduler to stabilize training during the first five epochs. We implemented an early stopping mechanism to interrupt the training process once the performance on the validation dataset stops improving after ten epochs for the MIMIC-III notes [99]. The tunable model hyperparameters (e.g., batch size, learning rate) are shown in Section 2.5.2.

2.4.7 Dataset: MIMIC-III

MIMIC-III is a publicly available database that contains de-identified EHRs (e.g., clinical notes, vital signs, laboratory measurements) for over 40,000 patients; more detailed information on the clinical aspects of the dataset can be found in Johnson et al. [69]. Each unique patient ID is associated with at least one hospital admission ID. Every hospital admission ID was annotated by trained human encoders with a subset

of the ICD-9 diagnosis and procedure codes. We formulated this problem as a multi-label document classification task. Therefore, our models aimed to predict a subset of labels, $y_i \in \{0,1\}^l$, associated with document i from the entire label set $|L|$ of the set of N documents $D = \{(x_i, y_i)\}_{i=1}^{N}$ by utilizing only the discharge summary note and potentially available addenda associated with the respective hospital admission ID. We performed experiments on the entire dataset for the full label set (MIMIC-III-Full) and on a smaller subset that contained documents associated with the 50 most frequent ICD-9 codes (MIMIC-III-50). We split the data into training, testing, and validation sets by following the splits proposed by Mullenbach et al. (2018) [99]. We tokenized each sequence, lower-cased all tokens, and removed non-alphabetic tokens or tokens that occurred less than three times in the corpus. We set the maximum document length to 3,000 tokens for the CNNs and RNNs and 4,096 tokens for the CLF to account for the differences in information content provided by word-tokens used by CNNs and RNNs and subword-tokens used by the CLF. Table 2.1 presents relevant statistics for this dataset.

2.5 Experiments

2.5.1 Evaluation Metrics

The model performance was assessed with evaluation metrics commonly featured in the document classification literature [143]. We report micro- and macro-averaged

Table 2.1: Statistics for MIMIC-III. $\overline{N_d}, (N_{d,\sigma})$ is the average number (standard deviation) of tokens per document, N_L is the total number of unique labels, N_{Cat} is the total number of categories, and $\overline{N_L}$ is the average number of labels per document.

Dataset	N_{total}	N_{train}	N_{val}	N_{test}	$\overline{N_d}(N_{d,\sigma})$	N_L	N_{Cat}	$\overline{N_L}$
MIMIC-III-Full	52,722	47,719	1,631	3,372	1,861 (949)	8,907	37	15.9
MIMIC-III-50	11,368	8,066	1,573	1,729	1,989 (968)	50	14	5.8

F1-scores, micro- and macro-averaged scores for the area-under-curve (AUC), and precision at k for k in [5, 8, 15] for the classification experiments.

2.5.2 Hyperparameter Optimization

We performed hyperparameter optimization for all text-encoder architectures on the validation split of MIMIC-III-50 by using a hill-climbing strategy to alleviate the computational burden; we utilized the found hyperparameters for the MIMIC-III-Full models. Table 2.2 lists and indicates the optimal hyperparameters for each architecture and dataset. The explored hyperparameters are based on previous work [49]. We selected a learning rate of $1e^{-4}$ for the CNN, BiGRU, and BiLSTM [49] and tuned the learning rate for the CLF with values provided by [84].

2.6 Results

2.6.1 Results: MIMIC-III – Single-head Label-Attention

Table 2.3 shows the performance scores for our single-head label-attention experiments on MIMIC-III-Full and MIMIC-III-50 and compares them with the performances of current benchmark models taken from the original studies. For MIMIC-III-Full, the results indicate that label-attention substantially improves classification performance over both baselines across all four text-encoder architectures. However, hierarchical random label-attention underperforms both baselines for the CNN- and CLF-based models. Models using pretrained reference information performed on average 0.014 points better than their randomly initialized counterparts on the macro-averaged F1-score across all text-encoder architectures. As expected, both RNNs are less sensitive in performance to different query matrix initializations than the CNN- and CLF-based models. Unexpectedly, incorporating code hierarchy via our hierarchical label-attention did not significantly boost model performance. Hierarchical pretrained label-attention performed better than hierarchical random

Table 2.2: For MIMIC-III we denote the selected hyperparameters for the architecture-specific methods for generating document embeddings as B, target-, and label-attention as T and L.

CNN	
Filter Size	100^{L}, 300, 500^{B}, 1000^{T}
Word/Label Embedding Dimension	$100^{B,L}$, 200, 300^{T}
Batch Size	$16^{T,L}$, 32^{B}, 64, 128
Dropout %	0, 0.15^{B}, 0.3^{T}, 0.5^{L}
BiGRU	
Hidden Size	128, 256^{L}, $512^{B,T}$
Word/Label Embedding Dimension	$100^{T,L}$, 200, 300^{B}
Batch Size	$16^{B,L}$, 32, 64, 128^{T}
Dropout %	0, 0.15^{T}, 0.3^{B}, 0.5^{L}
BiLSTM	
Hidden Size	128, 256, $512^{B,T,L}$
Word/Label Embedding Dimension	100^{B}, 200^{L}, 300^{T}
Batch Size	$16^{B,T,L}$, 32, 64, 128
Dropout %	0, $0.15^{B,T}$, 0.3, 0.5^{L}
CLF	
Word/Label Embedding Dimension	$100^{B,L}$, 200, 300^{T}
Batch Size	$4^{T,L}$, 6^{B}, 8, 16
Dropout %	0^{L}, 0.15, 0.3, $0.5^{B,T}$
Learning Rate	1e-$5^{B,T}$, 2e-5, 5e-5^{L}

Table 2.3: Performance results on MIMIC-III-Full and MIMIC-III-50 for the architecture-specific baseline (B), target-attention mechanism (T), random (R), pretrained (P), hierarchical random (HR), and hierarchical pretrained (HP) label-attention mechanism across multiple text-encoder architectures. We show performance results for previous and current state-of-the-art models directly taken from their work. Best overall performances are highlighted in bold; best performances of our models are bold and italicized.

Model	Attention Type	MIMIC-III-Full						MIMIC-III-50				
		AUC		F1		P@k		AUC		F1		P@k
		Macro	Micro	Macro	Micro	8	15	Macro	Micro	Macro	Micro	5
CAML [99]	R	0.895	0.986	0.088	0.539	0.709	0.561	0.875	0.909	0.532	0.614	0.609
DR-CAML [99]	P	0.897	0.985	0.086	0.529	0.690	0.548	0.884	0.916	0.576	0.633	0.618
MultiResCNN [82]	R	0.910	0.986	0.085	0.552	0.734	0.584	0.899	0.928	0.606	0.670	0.641
LAAT [143]	R	0.919	0.988	0.990	0.575	0.738	0.591	0.925	0.946	0.666	0.715	0.675
JointLAAT [143]	R	0.921	0.988	**0.107**	0.575	0.735	0.590	0.925	0.946	0.661	0.716	0.671
HyperCore [24]	P	**0.930**	**0.989**	0.090	0.551	0.722	0.579	0.895	0.929	0.609	0.663	0.632
D2SBERT [56]	R	—	—	—	—	—	—	—	—	0.629	0.686	—
EffectiveCAN [88]	R	0.921	**0.989**	0.105	0.581	0.755	0.604	0.920	0.945	0.668	0.717	0.664
HiLAT + ClinicalplusXLNet [87]	R	—	—	—	—	—	—	**0.927**	**0.950**	**0.690**	**0.735**	**0.681**
PLM-ICD* [61, 39]	R	0.926	**0.989**	0.104	**0.598**	**0.771**	**0.613**	0.917	0.938	0.663	0.705	0.657
CNN	B	0.827	0.977	0.021	0.379	0.619	0.476	0.836	0.888	0.454	0.558	0.571
	T	0.864	0.978	0.030	0.350	0.601	0.455	0.853	0.888	0.448	0.539	0.575
	R	0.862	0.979	0.039	0.463	0.642	0.491	0.891	0.922	0.565	0.640	0.628
	P	0.876	0.982	0.049	0.481	0.658	0.509	0.885	0.917	0.542	0.623	0.621
	HR	0.846	0.972	0.015	0.318	0.544	0.410	0.873	0.907	0.479	0.573	0.601
	HP	0.882	0.981	0.045	0.443	0.629	0.481	0.881	0.916	0.534	0.625	0.620
Bi-GRU	B	0.883	0.982	0.033	0.435	0.661	0.503	0.872	0.906	0.490	0.586	0.603
	T	0.883	0.982	0.038	0.438	0.647	0.493	0.855	0.892	0.452	0.552	0.581
	R	0.890	0.984	0.066	0.531	0.702	0.548	0.871	0.907	0.519	0.622	0.606
	P	0.887	0.984	0.068	0.535	0.706	0.551	0.879	0.915	0.552	0.643	0.624
	HR	0.898	0.984	0.063	0.526	0.699	0.545	0.870	0.906	0.526	0.614	0.608
	HP	0.903	0.985	0.068	0.533	0.705	0.550	0.880	0.916	0.568	0.654	0.626
Bi-LSTM	B	0.879	0.982	0.026	0.377	0.614	0.464	0.837	0.879	0.377	0.479	0.529
	T	0.876	0.981	0.032	0.381	0.599	0.451	0.849	0.883	0.450	0.540	0.560
	R	0.893	0.985	0.067	0.538	0.712	0.557	0.873	0.906	0.495	0.590	0.596
	P	0.892	0.984	0.066	0.538	0.711	0.555	0.880	0.914	0.546	0.634	0.618
	HR	0.901	0.985	0.061	0.522	0.699	0.545	0.866	0.901	0.494	0.584	0.594
	HP	0.911	0.986	0.067	0.537	0.709	0.555	0.880	0.914	0.546	0.634	0.618
CLF	B	0.867	0.977	0.025	0.383	0.612	0.462	0.840	0.879	0.437	0.541	0.565
	T	0.879	0.980	0.030	0.407	0.613	0.462	0.888	0.917	0.558	0.635	0.628
	R	0.903	0.985	0.057	0.524	0.724	0.571	0.898	0.925	0.575	0.644	0.640
	P	0.915	0.987	0.069	0.550	***0.741***	***0.589***	***0.905***	***0.930***	***0.597***	***0.664***	***0.646***
	HR	0.865	0.978	0.024	0.359	0.611	0.466	0.888	0.915	0.565	0.635	0.624
	HP	**0.930**	***0.988***	***0.070***	***0.552***	0.738	0.586	0.896	0.924	***0.597***	0.660	0.637

* Original authors [61] only provided results for MIMIC-III-Full; for completion showing reproduced results of MIMIC-III-50 [39].
Models: Convolutional Neural Network (CNN), bi-directional gated recurrent unit neural network (Bi-GRU), bi-directional long short-term memory neural network (Bi-LSTM), and transformer-based Clinical Longformer (CLF).

label-attention. Overall, CLF experienced the largest boost in performance with label-attention, followed by the Bi-LSTM and Bi-GRU, and the CNN benefited the least from label-attention. The CLF with hierarchical pretrained label-attention achieved the best performance across all our models with macro- and micro-averaged F1-scores of 0.070 and 0.552, respectively.

For MIMIC-III-50, the results suggest that random, pretrained, and hierarchical pretrained label-attention perform significantly better than the baseline methods across all text-encoder architectures except for the CLF-based model with random hierarchical label-attention. In contrast to MIMIC-III-Full, the initialization of the reference information for the smaller dataset can have a significant impact on model performance and suggests that selecting random or pretrained embeddings depending on the given text-encoder architecture optimizes performance. Context-based text-encoder architectures (e.g., Bi-GRU, Bi-LSTM, and CLF) achieved their best performances when using label-attention and hierarchical label-attention with pretrained embeddings. In contrast, CNN-based models had their best result when using random label-attention because of potential synergies with learning multiple n-gram patterns. Hierarchical pretrained label-attention achieved equal or better performance than pretrained label-attention across all text-encoder architectures, and the performance difference for the Bi-GRU is significant. Models that use hierarchical random label-attention performed worse than the remaining types of label-attention. Notably, the Bi-GRU performed significantly better than the Bi-LSTM for the RNN-specific baseline method by using the averaged hidden states over all time steps as the input for the classification layer.

Overall, CLF-based models achieved the highest performance, followed by the Bi-GRU, and last with similar performance the CNN and Bi-LSTM. The CLF-based model that uses pretrained label-attention scored the best performance with a macro- and micro-averaged F1-score of 0.597 and 0.664, respectively. The performance discrepancies between our best models and the current state-of-the-art model stem from the differences in the applied text-encoder architectures. For example, the

HiLAT [87] first splits each clinical document into smaller equally long chunks allowing it to use a transformer-based text-encoder architecture with dense self-attention to encode each chunk, whereas our best CLF-model is designed to encode the entire sequence at once using sparse self-attention. Sparse self-attention is more likely to underperform compared to dense self-attention but provides computational benefits like memory usage. By reducing the effective document, the HiLAT is able to learn more fine-grained semantic information from the clinical document, whereas, our CLF-based model makes long-range connections, potentially neglecting local semantic information. We re-emphasize that for the purpose of this study, we focus on more fundamental architectures, which is why we chose the more straightforward CLF as our Transformer baseline rather than a more complicated architecture such as the HiLAT.

For MIMIC-III-Full, we counted the number of times a specific type of label-attention achieved the highest performance for all labels in the label space. We then examined how the strategy used to initialize label-attention relates to the text-encoder architecture. Table 2.4 summarizes the retrieved counts per frequency quartile for all labels for which at least one attention type achieved a nonzero performance; no label in Q1 achieved nonzero performance across all evaluated models, and thus Q1 is omitted. The results indicate that (i) all models experience difficulties in predicting minority classes (e.g., only 1% of labels in Q2 were predicted correctly), (ii) different text-encoder architecture and label-attention type combinations can learn to correctly classify a different subset of labels, (iii) hierarchical pretrained label-attention helps CNN- and CLF-based models classify minority classes better (e.g., Q2, Q3), and (iv) the majority of labels achieve their best performance with models utilizing either hierarchical pretrained or pretrained label-attention mechanisms.

Table 2.4: Frequency count of labels that achieve their best performance with either random (R), pretrained (P), hierarchical random (HR), or hierarchical pretrained (HP) label-attention mechanisms broken down by frequency quartile across all text-encoder architectures for MIMIC-III-Full. The counts only consider labels for which at least one attention type achieved a nonzero performance across all models; no label in Q1 achieved nonzero performance, and thus Q1 is omitted. Different model combinations may correctly classify a different subset of labels.

Encoder	Label Attention	Label Frequency Count Q2	Q3	Q4	Total
CNN	R	2	19	217	238
	P	2	79	629	710
	HR	0	0	5	5
	HP	10	166	424	600
Bi-GRU	R	5	74	364	443
	P	8	79	451	538
	HR	4	89	289	382
	HP	5	65	376	446
Bi-LSTM	R	7	84	433	524
	P	6	81	459	546
	HR	2	52	198	252
	HP	0	86	429	515
CLF	R	0	55	189	244
	P	1	85	641	727
	HR	0	5	17	22
	HP	0	98	639	737
Number of Correctly Predicted Labels:		30	452	1,752	2,234
Total Number of Predicted Labels:		2,437	2,312	2,248	8,907
Label Frequency Range:		2–5	6–27	28–20,046	

2.6.2 Results: MIMIC-III – Multi-head Label-Attention

We performed extensive experiments on a multi-head version of the label-attention mechanism. Table 2.5 lists the micro- and macro-averaged F1-scores of our experiments for both MIMIC-III datasets. For both datasets, all text-encoder architectures that used multi-head label-attention experienced significant performance improvements over their single-head versions. Significant performance improvements were predominantly achieved by using multi-head attention with either two or four attention heads. Additionally, hierarchical random label-attention experienced the strongest performance increase from using multi-head attention compared with the remaining types of label-attention. We believe this significant performance increase stems from each attention head attending to its own set of features of the input, thereby allowing the model to learn more important parts more easily.

For MIMIC-III-Full, multi-head label-attention significantly improved the macro-average scores across all text-encoder architectures with the strongest performance of 0.092 achieved by the Bi-GRU using random multi-head label-attention. This indicates that multi-head label-attention can improve the classification of minority classes, which is beneficial for the identification of rare diseases. Further, the micro-averaged F1-scores for the Bi-GRU-based models were unaffected by the implementation of multi-head attention, whereas the Bi-LSTM-, CNN-, and CLF-based models experienced significant performance improvements. The CLF-based model that used hierarchical pretrained multi-head label-attention performed slightly better than the single-head version with a micro-averaged score of 0.554. For MIMIC-III-50, the Bi-LSTM benefited tremendously from using multi-head attention and showed significant improvements over the single-head version for random, hierarchical random, and hierarchical-pretrained label-attention across all numbers of attention heads. Multi-head attention had less impact on the performance of CLF-based models compared with the remaining three encoders. Our results suggest that the ideal

Table 2.5: Multi-head label-attention results on MIMIC-III. Best performance across all models is highlighted in bold, * indicates improved performance over the single-head label-attention counterpart; ** indicates significant improvement.

Dataset	Encoder	Heads	Random		Pretrained		Hierarchical Random		Hierarchical Pretrained	
			Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
MIMIC-III-Full	CNN	2	0.087**	0.510**	0.082**	0.522**	0.079**	0.509**	0.081**	0.516**
		4	0.004	0.187	0.078**	0.522**	0.077**	0.509**	0.081**	0.514**
		6	0.027	0.313	0.027	0.394	0.072**	0.500**	0.070*	0.510**
	Bi-GRU	2	**0.092****	0.522	0.063	0.506	0.087**	0.520	0.088**	0.521
		4	0.087*	0.523	0.073*	0.530	0.077**	0.509	0.080**	0.518
		8	0.083*	0.517	0.072*	0.526	0.064*	0.509	0.074**	0.522
	Bi-LSTM	2	0.080**	0.542*	0.079**	0.548**	0.087**	**0.554****	0.081**	0.544*
		4	0.082**	0.540*	0.068*	0.539*	0.084**	0.547**	0.074*	0.540*
		8	0.081*	0.535	0.068*	0.534	0.064*	0.526*	0.068*	0.532
	CLF	2	0.064*	0.535*	0.066	0.548	0.050**	0.513**	0.072*	**0.554***
		4	0.063*	0.534*	0.066	0.548	0.065**	0.535**	0.071*	0.553*
		8	0.035	0.460	0.051	0.523	0.072**	0.547**	0.075*	0.544
MIMIC-III-50	CNN	2	0.566*	0.646*	0.556**	0.638*	0.556**	0.639**	0.550**	0.634**
		4	0.558	0.638	0.549*	0.631*	0.563**	0.641**	0.542*	0.628*
		6	0.553	0.636	0.542	0.626*	0.537**	0.630**	0.528	0.621
		10	0.545	0.631	0.534	0.620	0.539**	0.630**	0.536*	0.625
	Bi-GRU	2	0.559**	0.637**	0.547	0.635	0.581**	0.650**	0.567	0.645
		4	0.550**	0.638**	0.554*	0.643	0.580**	0.646**	0.577*	0.652
		8	0.530*	0.620	0.565*	0.645*	0.573**	0.644**	0.562	0.646
		16	0.521*	0.612	0.554*	0.639	0.555**	0.635**	0.555	0.638
	Bi-LSTM	2	0.558**	0.642**	0.549*	0.640	0.558**	0.646**	0.563**	0.651**
		4	0.542**	0.628**	0.564**	0.649**	0.564**	0.647**	0.557**	0.645**
		8	0.543**	0.630**	0.559*	0.644*	0.553**	0.640**	0.553**	0.644**
		16	0.534**	0.617**	0.543	0.635*	0.543**	0.627**	0.547*	0.638*
	CLF	2	0.580*	0.645*	0.592	0.664	0.575*	0.640*	**0.600***	**0.669***
		4	0.569	0.643	0.588	0.656	0.582*	0.649*	0.593*	0.661*
		8	0.561	0.640	0.585	0.651	0.573*	0.639*	0.579	0.653
		16	0.557	0.631	0.582	0.652	0.570*	0.638*	0.571	0.646

number of attention heads depends on the architecture used, as well as trade-off considerations between improvement gains and compute requirements.

Our results on MIMIC-III revealed the following: (i) label-attention generally improves classification performance over architecture-specific baselines and target-attention methods, (ii) using label-attention with pretrained embeddings is preferable over random embeddings for contextual-based text-encoders (e.g., RNNs and transformers), (iii) using the transformer-based CLF achieves the best performance at the cost of a significant increase in compute time (i.e., GPU-usage), (iv) the Bi-GRU outperforms the Bi-LSTM on aggregate, (v) incorporating code hierarchy via label-attention can significantly improve model performance for biomedical text classification tasks and for the Bi-GRU and CLF in particular, and (vi) multi-head label-attention has a strong impact on the classification performance, of the CNNs and RNNs but only a marginal affect on CLF-based models.

2.7 Discussion

2.7.1 Comparison of Attention and Energy Scores

Modern attention mechanisms consist of two primary operations: (i) the matrix multiplication between the input sequence in K and the reference information in Q, which results in the energy scores, and (ii) the softmax operation, which produces the attention scores by normalizing the raw energy scores. Both energy and attention scores are essential to direct a model's focus to the most relevant tokens associated with a medical concept; however, both sets of scores differ in their interpretation depending on the nature of the desired analysis.

The primary difference between energy scores and attention scores lies in the interpretation of the output of the deployed mathematical operations. The matrix multiplication allows us to interpret the magnitude and sign of the output as the

strength and direction of the association between the input in K and the label-specific reference information in Q. In this way, the output, or energy score, provides a global view of how related the token is to a medical concept regardless of its current document or corpus. In contrast, the softmax function output can be interpreted as pseudo-probabilities that indicate the local importance of a token for predicting a positive association relative to all other words in a document. The final attention score is therefore influenced by the relevance of all other tokens and the total length of a document [117]. While attention scores help the model discriminate between signal and noise in a specific document, the softmax operation retrieves the underlying global association between a token and the label-specific reference information initially captured by the energy score and, ultimately, with the associated medical concept.

Consider the following two energy score sequences with three tokens $E1 = [-1.0, -2.0, -3.0]$ and $E2 = [3.0, 2.0, 1.0]$. After applying the softmax operation, we get identical attention scores $A1 = A2 = [0.665, 0.245, 0.090]$ for both energy score sequences despite having different initial energy scores. This example illustrates that a token may appear relevant locally within the context of a document but may possess a negative underlying global association with the medical concept that the token is trying to explain. This ambiguity is critical because the medical profession's acceptance of neural networks as a diagnostic tool depends on the traceability of a model's reasoning for its predictions [98].

Lastly, our work explores how different initialization strategies of the reference information in Q impact model performance and a model's motivation behind its predictions. These differences in Q across the label-attention types can be directly observed in the generated energy scores but not in the attention scores, making energy scores the ideal candidate for understanding the effect of different attention architectures and initialization strategies. However, we want to emphasize that in clinical practice the combination of both sets of scores are essential for sufficiently interpreting a model's predictions.

2.7.2 Random vs. Pretrained Label Reference Information

This study investigates the impact of an attention mechanism's initialization strategy on classification performance and potential differences between the initialization of an attention mechanism's query matrix with random or pretrained embeddings across various text-encoder architectures. Our results in Section 2.6 indicate observable differences between both strategies for smaller datasets such as MIMIC-III-50, whereas the differences become negligible for larger datasets.

Randomly initialized models try to learn label-specific reference information from the available samples during training. This method relies on the training corpus having a diverse set of documents to provide information about the entire label space; however, the label spaces of clinical text classification tasks are often dominated by a small but frequently occurring set of labels which inhibits the learning of representative features for infrequent labels. This imbalanced representation of learned features can sometimes be addressed by incorporating prior knowledge about the label space, such as the initialization of an attention mechanism's query matrix with pretrained embeddings of textual code descriptions. The purpose of using these pretrained embeddings is to prime the model with information relevant to each label and to push model performance. For example, if a document has the ICD-9 code *427.31 - Atrial Fibrillation* assigned to it, then we expect the most relevant tokens or phrases to be similar to *atrial*, *fibrillation*, or a combination of both. We think the encodings of relevant tokens will be closer to the label-specific query vector in the embedding space, thereby resulting in larger energy scores. Thus, we hypothesize that the choice of text-encoder architecture and type of label-attention directly affect the resulting energy scores on a token level. To investigate this hypothesis, we retrieved the energy score distributions from the test corpus of MIMIC-III-50 for all label-attention types across the four text-encoder architectures (Fig. 2.3).

We can infer from Fig. 2.3 the following differences between the energy score distributions for the label-attention mechanisms: (i) RNN- and CLF-based models are

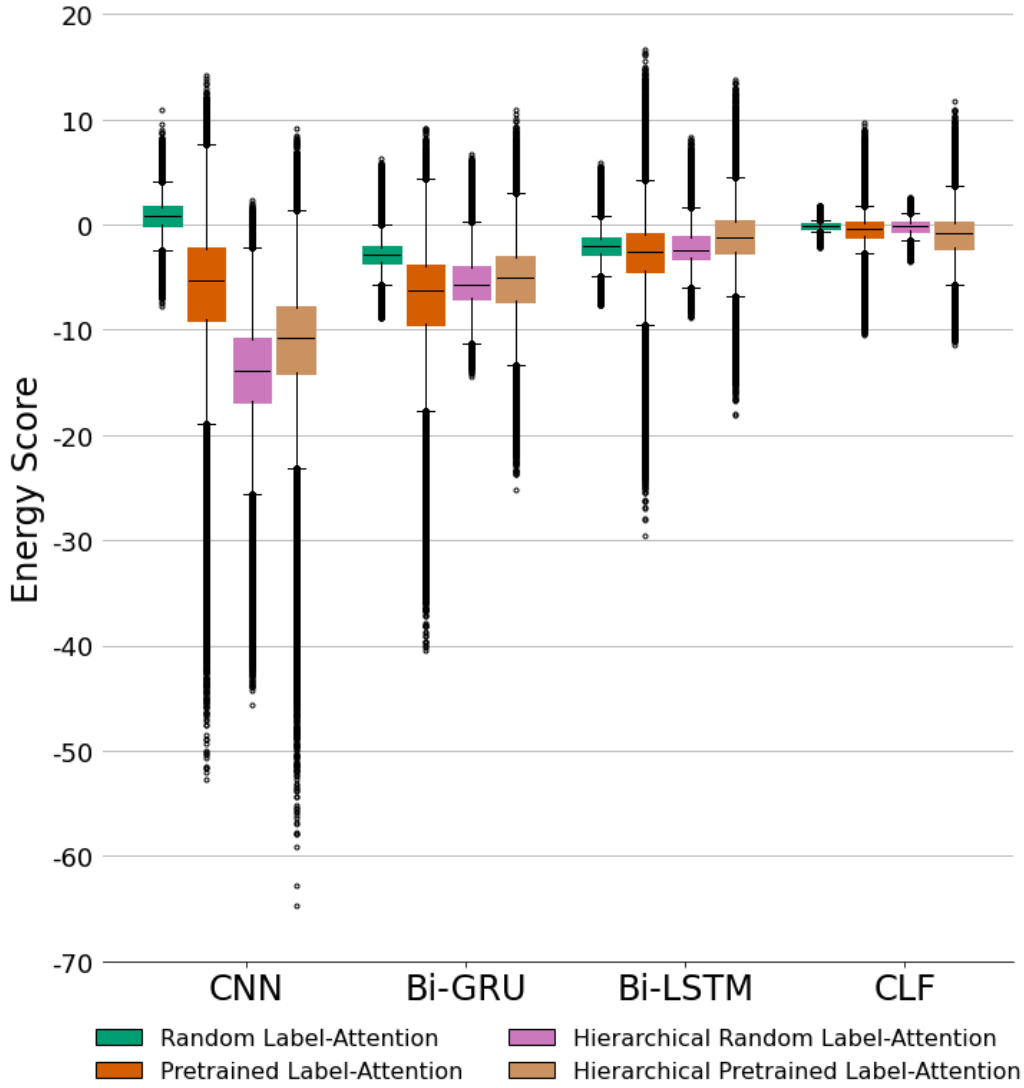


Figure 2.3: Differences between the energy score distributions for all types of label-attention and text-encoder architectures extracted from the MIMIC-III-50 test documents. The dots are outliers and represent token-specific energy scores that lie beyond $\pm 1.5IQR$.

less sensitive in their energy score distributions to differently initialized query matrices (i.e., reference information) compared with the CNN-based model. We think using the contextual-based encoding of clinical documents allows RNNs and transformers to generalize better and produce more similar token-vector representations, and this increases robustness for the computation of the energy scores. In contrast, CNNs learn embeddings from multiple n-grams disconnected from the context of the document, and this results in a more diverse set of token-vector representations that affect the computation of the energy scores. (ii) The majority of energy scores are negative across all label-attention and text-encoder architecture combinations, except for CNN when using random label-attention. This is reasonable because only a minority of tokens are indicative of medical classifications in clinical text classification tasks. (iii) Label-attention mechanisms with pretrained embeddings have larger extreme values and a larger spread for their energy scores compared with their respective randomly initialized counterparts. This is an artifact of using the dot product between a token representation and the embedded, label-specific reference information used to calculate the energy score for each token. Essentially, the model can more confidently distinguish between relevant tokens and noise, thereby improving its discriminatory ability.

As mentioned above, prior knowledge can be used to address the label imbalance issues commonly found in clinical text classification tasks to improve the prediction performance of rare classes. We retrieved the model performance broken down by label frequency quartile to investigate if pretrained embeddings can improve prediction performance on minority labels for MIMIC-III-Full. The results in Table 2.6 shows that models overall and across all text-encoder architectures have improved performance on minority labels (i.e., Q2 and Q3) when using hierarchical pretrained or pretrained label-attention compared against their respective randomly initialized counterpart. This indicates that using pretrained embeddings in an attention mechanism helps the model learn meaningful features used to search for tokens or phrases relevant to the model.

Table 2.6: Model performance per quartile with random (R), pretrained (P), hierarchical random (HR), or hierarchical pretrained (HP) label-attention organized by frequency quartile averaged and across all text-encoder architectures for MIMIC-III-Full. No model was able to predict labels in Q1, and thus Q1 is omitted.

Encoder	Label Attention	Micro F1-Score Q2	Q3	Q4
Average	R	0.004	0.048	0.532
	P	0.006	0.058	0.545
	HR	0.002	0.030	0.447
	HP	0.005	0.061	0.536
CNN	R	0.002	0.023	0.485
	P	0.002	0.049	0.504
	HR	0	0.002	0.331
	HP	0.006	0.061	0.471
Bi-GRU	R	0.005	0.068	0.547
	P	0.012	0.072	0.552
	HR	0.002	0.058	0.543
	HP	0.011	0.072	0.550
Bi-LSTM	R	0.008	0.062	0.555
	P	0.009	0.060	0.555
	HR	0.004	0.055	0.539
	HP	0.004	0.060	0.554
CLF	R	0	0.037	0.541
	P	0.001	0.050	0.567
	HR	0	0.006	0.373
	HP	0	0.054	0.569

2.7.3 Model Interpretability: Phrase-Level Evaluation

Showcasing a model's ability to provide interpretability helps build trust in clinicians in AI-driven clinical support systems and improve patient outcomes [5]. To investigate the general capability of models that use label-attention to provide such evidence and potential differences between the various label-attention types on a global level, we extracted the most relevant phrase per code across all MIMIC-III-50 test documents. Specifically, for each model, we extracted the key phrase, kp, with the largest average energy score out of a pool of phrases with different sizes from the entire set of label-specific energy score sequences, E,

(2.8):

$$kp = max(\frac{1}{s}\sum_{i=j}^{j+s-1} e_i) \qquad i : [0, N], \tag{2.8}$$

where e_i is the energy score of the i^{th} token in the sequence with N tokens, and s is the phrase size with either (i) three, four, or five tokens for the RNNs/CNNs or (ii) four, six, or eight tokens for the CLFs. The phrase size differences account for the reduced information provided by the CLFs' subword-tokens.

Table 2.7 provides examples for extracted text snippets containing highly relevant phrases including their word-specific energy scores for the prediction of ICD-9 codes, e.g., *lactic acidosis* for the ICD-9 code *276.2 Acidosis*. A clinician can use the energy scores in tandem with the respective attention scores to interpret the prediction of a medical code by answering the following two questions: i) How strong is the global relationship between a word and the medical concept? ii) How important is the same word relative to all other words in the clinical document? The first question can be answered using the energy scores where high energy scores indicate that particular word is universally associated with a medical concept across all documents in the corpus, e.g., the model is able to associate *hypoxic*, a term describing the deficiency of oxygen in tissue, with the ICD-9 code *518.81 Acute Respiratory Failure*. The latter question can be answered using the attention scores where larger scores represent

higher importance within that particular document. See Section 2.7.1 for a more detailed discussion on the differences between energy and attention scores.

A medical professional evaluated a total of 256 extracted phrases for their associations (3 - Strong Association, 2 - Somewhat Associated, and 1 - No Association) with the respective ICD-9 codes. The results of this evaluation are shown in Table 2.8 and indicate differences in evidence-based interpretability between the text-encoder architectures and between the label-attention mechanisms. On the text-encoder architecture level, we see that context-based text-encoder architectures are more likely to provide immediate evidence across a larger subset of ICD-9 codes compared with the n-gram-based CNN; this difference is more prominent for infrequent codes associated with Q1. We think that the CNNs' design causes the model to put emphasis on phrases that help the model to better discriminate between the codes but are not meaningful to humans. For the CNN, Bi-GRU, and the CLF, the pretrained label-attention mechanisms are more likely to provide evidence for a larger set of medical codes compared with their respective randomly initialized counterparts. We think that initializing attention mechanisms with pretrained embeddings helps the

Table 2.7: Model interpretability with label-attention for two ICD-9 codes. The examples text (T) snippets were taken from two evaluated test documents using the BiGRU with hierarchical pretrained label-attention; phrases marked as important are highlighted in bold. The numbers above each word represent their respective energy (E) or attention (A) scores.

276.2: *Acidosis*

E:	1.02	1.30	1.20	2.39	3.04	4.81	6.32	
A:	0.0008	0.0011	0.0010	0.0033	0.0063	0.0372	0.1687	
T:	... *patient*	*continued*	*to*	*have*	*raising*	***lactic***	***acidosis.***	

E:	4.40	3.35	2.38	1.97				
A:	0.0246	0.0086	0.0033	0.0022				
T:	*PT*	*was*	*ultimately*	*found*	...			

518.81: *Acute Respiratory Failure*

E:	3.69	3.85	4.19	5.43	6.35	6.61	5.95		
A:	0.0029	0.0034	0.0048	0.0168	0.0419	0.0544	0.0280		
T:	... *on*	*floors*	*for*	***hypoxic***	***respiratory***	***failure.***	*This*		

E:	6.08	6.20	5.37	4.46	3.49	2.65	2.77		
A:	0.0321	0.0361	0.0157	0.0063	0.0024	0.0010	0.0012		
T:	***respiratory***	***failure***	***was***	*quickly*	*reversed*	*with*	*diuresis*	...	

Table 2.8: Expert evaluation of the most-relevant label-specific key phrase per frequency quartile for the random (R), pretrained (P), hierarchical-random (HR), and hierarchical-pretrained (HP) label-attention mechanisms. Results are presented across four text-encoder architectures. A ✓indicates a strong association of a key phrase with an ICD-9 code as deemed by the medical expert.

		ICD-9 Codes*																	
		Q1					Q2					Q3					Q4		
Encoder	Attention	93.90	45.13	V15.82	287.5	507.0	276.1	V58.61	36.15	38.91	276.2	530.81	39.61	599.0	518.81	38.93	401.9	Total
		Label Frequency →																
CNN	R	✓							✓			✓	✓	✓	✓		✓	7
	P				✓		✓		✓	✓	✓		✓	✓		✓		8
	HR				✓		✓		✓	✓	✓		✓	✓		✓		8
	HP				✓		✓		✓	✓	✓		✓	✓		✓	✓	9
Bi-GRU	R		✓		✓		✓	✓	✓	✓	✓		✓	✓				9
	P		✓		✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	11
	HR			✓			✓		✓		✓		✓	✓	✓		✓	8
	HP		✓		✓	✓	✓	✓	✓		✓		✓	✓			✓	10
Bi-LSTM	R			✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓		11
	P				✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	10
	HR			✓	✓		✓		✓	✓	✓	✓	✓	✓		✓		10
	HP				✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	10
CLF	R	✓							✓			✓	✓	✓	✓	✓	✓	8
	P	✓		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13
	HR				✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	10
	HP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	15

*The top two most- and least-frequent ICD-9 codes per frequency quartile were extracted from the MIMIC-III-50 test corpus.

model look for phrases similar to the code description. For example, code 36.15 *Single internal mammary-coronary artery bypass* models often assigned the highest importance to phrases such as *coronary artery disease* for the CNN or *coronary artery bypass graft* for the CLF. However, the results indicate that both pretrained label-attention mechanisms assign high relevance to specific phrases across all labels. For example, the phrase *potassium chloride meq* was indicated as highly relevant across all labels while only being positively associated with a subset of codes. The retrieval of the same phrase as important across labels with opposing associations may indicate that label-attention is prone to propagating similar reference information across the different label-specific queries.

2.7.4 Attention to Code Hierarchy

This study proposes an extension of label-attention that can incorporate the hierarchical structure of a medical encoding system (e.g., ICD-9) without needing to train an auxiliary network. Contemporary work often deploys auxiliary networks (e.g., GNNs) to learn the code hierarchy and infuse the primary model with information on the hierarchical relationships between the labels to improve predictions [29]. In contrast, our hierarchical label-attention method incorporates hierarchical relationships by learning a set of features associated with high-level categories for each document that are then propagated downstream to enrich the low-level label reference information (i.e., each label query is enriched with the features of its parent category). We hypothesize that our category-wise incorporation of hierarchical information will lead to performance improvement on a category level. To test this hypothesis, we retrieved the label-specific performance and computed an aggregated, macro-averaged F1-score for all 14 categories associated with the MIMIC-III-50 classification task. The results are shown in Fig. 2.4. For the CNN, Bi-GRU, and CLF, the majority of categories exhibit improved performance. In particular, the Bi-GRU- and CLF-based models strongly benefited from deploying hierarchical label-attention. We showed

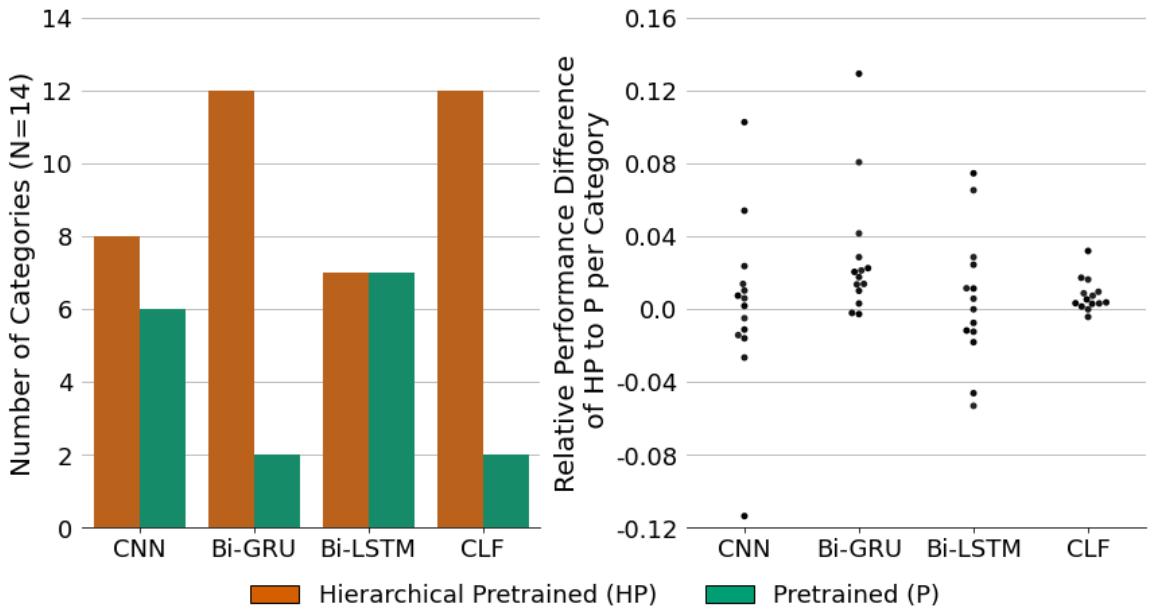


Figure 2.4: Achievable performance improvements per category of hierarchical pretrained label-attention (HP) over pretrained label-attention (P) for CNN, Bi-GRU, Bi-LSTM, and CLF. Aggregated number of categories achieving their best performance with hierarchical pretrained or pretrained label-attention, respectively (left). The relative performance difference is computed by subtracting P from HP (i.e., each point above 0 represents cases in which HP performs better. Points below 0 represent cases in which P performs better). Each statistic was computed as the aggregated, macro-averaged F1-score of all associated labels with their parent category (right).

that enriching the label-specific reference information with high-level categorical information leads to improved classification performance overall (see Section 2.6) and on a category level. The analysis also demonstrates that the performance gains are equally distributed across all categories.

2.8 Conclusion

In this work, we compared two strategies to initialize the reference information of the label-attention's query matrix across four text-encoder architectures, CNN, Bi-GRU, Bi-LSTM, and the transformer-based CLF, on the two common MIMIC-III clinical text classification tasks. We compared these methods against common methods to generate document embeddings and against target-attention. We showed that utilizing pretrained reference information in an attention mechanism's query matrix improves classification performance across smaller label spaces, but the performance increase is negligible with an increasing number of labels and documents. Additionally, we demonstrated that performance improvements are achievable by incorporating the hierarchical structure of medical encoding systems via the attention mechanism without requiring a second external network. Furthermore, we showed that label-attention mechanisms can provide highly relevant phrases allowing clinicians to interpret the prediction of a specific medical code. We expect our research to be helpful for future work that explores label-attention and pretrained reference information to maximize the potential of deep-NLP models for the automated classification of clinical documents with medical codes. The source code is available at https://github.com/cmetzner93/attention_mechanisms.

Chapter 3

Deformable phrase level attention: a flexible approach for improving AI based medical coding

Disclosure Statement

A version of this chapter was submitted to the Journal of Artificial Intelligence in Medicine; the submitted manuscript has been under review since then. A preliminary version has been published at the Social Science Research Network (SSRN).

Metzner, Christoph Simon and Gao, Shang and Herrmannova, Drahomira and Gounley, John and Hanson, Heidi, Deformable Phrase Level Attention: A Flexible Approach for Improving Ai Based Medical Coding. Available at SSRN: https://ssrn.com/abstract=4864687 or http://dx.doi.org/10.2139/ssrn.4864687

Author's contributions are as follows: Christoph Metzner designed and executed the study; Shang Gao, Drahomira Herrmannova, and Heidi Hanson contributed to the experimental design and analysis of the results. CM organized, programmed, and performed all experiments; John Gounley helped organizing the experiments. CM wrote the manuscript. All authors read, reviewed, and approved the final manuscript.

No revisions to this chapter have been made since the original submission.

Abstract

Objective: Improving the AI-driven automated medical encoding of clinical text plays a vital role in gathering information on the occurrence of diseases to improve population-level health. This work presents a novel attention mechanism designed to enhance text classification models and ensure appropriate classification of medical concepts in unstructured electronic health records.

Materials and Methods: We developed a deformable, phrase-level attention mechanism to identify important lexical word-level and contextual phrase-level information from clinical text documents. We evaluated conventional and transformer-based deep learning models that we extended with our attention mechanism on the extraction of critical cancer information (e.g., site, subsite, laterality, histology,

behavior) from 629,908 electronic pathology reports and on the automated medical encoding of 52,722 hospital discharge summaries.

Results: Transformer-based models with the deformable, phrase-level attention mechanism achieved the best performance on the extraction of critical cancer information from pathology reports. Conventional- and transformer-based models show similar or better performance than their baseline counterparts on the automated medical encoding of clinical documents.

Discussion: The addition of phrase-level information allowed models extended with our proposed method to outperform standard word-level attention. Our method showed favorable properties for the real-world application in terms of model robustness and phenotyping. These results indicate that our method is promising for automated data harmonization for common data models.

Conclusion: This work proposes a novel deformable, phrase-level attention mechanism that enhances text classification models in the extraction of medical concepts from clinical text documents. We demonstrate strong performances on two clinical text datasets and showcase real-world deployability of our method.

3.1 Background and Significance

AI is revolutionizing the technological landscape across all domains, including medicine [120]. For example, AI-driven clinical support systems demonstrate increased efficiency and accuracy of early detection and diagnosis methods such as medical imaging analysis (e.g., MRIs, CT scans) [141]. Additionally, these AI-driven systems can be helpful in personalized medicine (e.g., genomic sequencing) [114] and are vital in automated medical encoding of massive amounts of clinical text [50, 87]. Efficient automated medical encoding of unstructured electronic health records (EHRs) can help with the harmonization and curation of clinical data

elements for common data models (CDMs) such as the Minimal Common Oncology Data Elements (mCODE) initiative [106] or the Fast Healthcare Interoperability Resources (FHIR) standards [10]. Transitioning from manual to automated curation of CDMs can help with the near real–time detection of current cancer trends among the American population [109, 41] or near real–time detection of disease during a biopreparedness event [72, 126]. However, as the sheer quantity of clinical text has been steadily increasing [21, 130], the manual encoding of clinical documents is becoming prohibitively difficult because it is time-consuming, cost-intensive, error prone, and demands significant expertise [135]. Therefore, to improve population-level health through the automated data harmonization and curation of CDMs, we designed an AI-based system that efficiently encodes unstructured EHRs with medical codes by using natural language processing (NLP) and ML.

NLP and ML algorithms have been investigated as a promising approach to automate medical encoding of clinical text [47, 139]. Early approaches used rule-based NLP [47, 139] and statistical ML algorithms [147, 101] to extract cancer information from pathology reports, but unmanageable complexity in handwritten rules and manual feature engineering [123] necessitated a shift towards deep neural networks capable of learning salient semantic and contextual features directly from the input data. To help these models focus on distinct parts of the input and improve the classification of clinical text, state-of-the-art text classification models started to deploy attention mechanisms [99, 50]. Attention mechanisms enable a model to focus on important global relationships between individual words by using internally generated soft weights. However, these word-level attention mechanisms often overlook crucial local semantic context at the phrase level [85].

Phrases are conceptual units that consist of multiple words and provide additional local context that helps define a word's full meaning [150]. Linguistically, the meaning of a word is defined by itself and its context [42]. Since traditional attention mechanisms neglect local contextual information, previous work proposed methods to overcome this shortcoming at the phrase level. For example, previous work focused

on extracting only the phrase-level context, thereby ignoring the lexical meaning of a word [103, 92] or requiring prior syntactical parsing through the text to identify correct phrases [149], or on determining the context by using a fixed-size moving window [128], thereby disregarding that different words may need varying context sizes [92]. Therefore, we hypothesize that utilizing an attention mechanism that can accurately extract salient lexical and flexible-sized context information from clinical text will lead to improved model performance. In particular, we think that extending a model's visibility from the word level to its surrounding context will help with the classification of medical information, which is linguistically complex. For example, in the case of *non-small cell carcinoma* (8046), a histology code described with four words, a word-level attention mechanism might focus on individual words such as *carcinoma* and possibly cause confusion with the histology type *adenocarcinoma* (8140), whereas a phrase-level attention mechanism can dynamically expand its scope to correctly identify the class by considering the broader context of the phrase.

3.2 Objective

In this work, we propose a novel attention mechanism called deformable phrase-level attention (DPLA) to dynamically capture the lexical and contextual semantic information of each word to improve the automated classification of clinical documents. The DPLA functions as a distinct module by extending any deep learning–based architecture to enhance the extraction of key features from a given encoded document. We performed two sets of experiments on multi-class and multi-label clinical text datasets. The first set of experiments focused on improving the automated information extraction* of critical cancer data elements (e.g., site, subsite, laterality, histology, behavior) on a multi-class dataset that contains electronic pathology reports collected by the National Cancer Institute's (NCI) Surveillance,

*In traditional NLP terms, we define this task as a text classification; however, we loosely refer to the task as information extraction because the goal is to retrieve cancer elements from electronic pathology reports.

Epidemiology, and End Results (SEER) registries between 2004 and 2022 [2]. Our proposed model, the transformer-based Clinical Longformer (CLF) [84] extended with the proposed DPLA, significantly outperforms all baselines and achieved the highest performance across all oncological tasks. The second set of experiments was performed on the publicly available MIMIC-III hospital discharge summary notes [69] for an important and widely accepted task that involves the classification of multiple medical concepts from a single document. We show that extending three commonly deployed text-encoder architectures—convolutional neural network (CNN) [74], bi-directional gated recurrent unit neural network (Bi-GRU) [31], and the CLF—with the DPLA generally results in higher performance than the same text encoders with the standard label-wise attention mechanism [99]. Finally, we discuss the suitability of the DPLA in real-world scenarios such as harmonization of critical phenotype information found in clinical records.

3.3 Materials and Methods

3.3.1 Datasets

Multi-Class Dataset: SEER Electronic Pathology Reports

We retrieved a total of 629,908 electronic pathology reports from the NCI SEER [2] cancer registries in Kentucky, Louisiana, New Jersey, Washington, Utah, and New Mexico to assemble a representative cross-section of cancer occurrence in the American population for 69 primary cancer sites (see Table A.2). We performed in-distribution (ID) experiments ($n = 523,411$) with data from the first five registries and out-of-distributions (OOD) experiments ($n = 106,497$) on New Mexico's reports. Every year certified cancer registrars (CTRs) manually annotate millions of electronic pathology reports with ground-truth labels of important cancer information, including primary cancer site, subsite, laterality, histology, and behavior, by following North American Association of Central Cancer Registries standards [3]. We followed the

preprocessing procedure described by Gao et al. [49]. For more information, including partitioning of both datasets, see Appendix A.1.

Multi-Label Dataset: MIMIC-III Hospital Discharge Summary Notes

The second dataset contains deidentified hospital discharge summary notes from the publicly available MIMIC-III database [69]. Trained human encoders manually annotated each note with International Statistical Classification of Diseases, Ninth Revision (ICD-9) codes. For our experiments, we utilized two widely accepted subsets: (1) MIMIC-III-Full ($n = 52,722$), which contains all documents with the full label set, and (2) MIMIC-III-50 ($n = 11,368$), which comprises notes annotated with the 50 most frequent labels. We follow the training, validation, and testing splits provided by Mullenbach et al. [99]. Each clinical note was tokenized, lower-cased, and cleaned from non-alphabetic and rare (i.e., occurring less than three times in the corpus) tokens.

3.3.2 General Model Architecture

The structure of the evaluated clinical text classification models is shown in Figure 3.1 and consists of (i) a word-embedding layer that takes a text sequence $X = [x_1, x_2, \ldots, x_N]$ with N integer-indexed tokens as the input and maps each token to its associated word vector representation with size d_e creating a token-embedding matrix $X_E \in \mathbb{R}^{N \times d_e}$. X_E is then passed to (ii) a text-encoder architecture to learn a latent document representation $H \in \mathbb{R}^{N \times d_h}$ with hidden dimension d_h. The latent document representation H is then passed into (iii) an attention mechanism to extract the most important features from H; this results in the final document vector matrix $C \in \mathbb{R}^{m \times d_h}$ where m depends on the classification task and selection of (iii) (see Section 3.3.3). C is then passed to a (iv) linear classification layer with projection weights $W \in \mathbb{R}^{|L| \times d_h}$ and bias $B \in \mathbb{R}^{|L| \times 1}$—$|L|$ represents the number of classes for a given task—followed by (v) an activation function to generate the class-specific

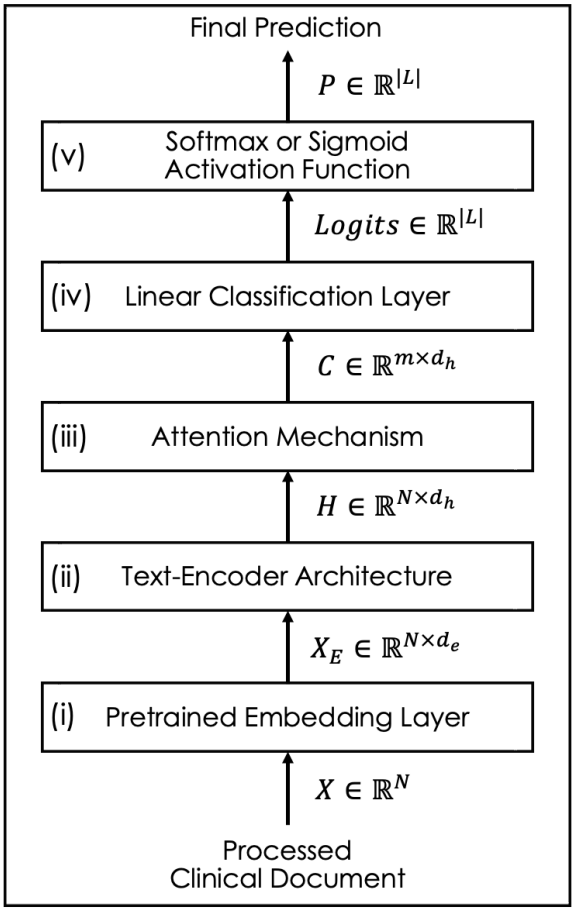


Figure 3.1: General model architecture. Each model takes an integer-indexed clinical text document with N tokens as the input and embeds each word to its corresponding vector representation by using (i) a pretrained word-embedding layer. A (ii) text-encoder architecture then encodes each word in X_E before passing the document H to either (iii) an attention mechanism to extract the most important features from the text sequence to generate the final document vector representation C. C is than passed to (iv) a classification layer paired with (v) an activation function to generate the class-specific probabilities. Component (i) depends on the selection for (ii), and (v) depends on the type of classification task (i.e., multi-class – softmax vs. multi-label – sigmoid). The variable m indicates the number of elements in the output matrix and depends on the type of the classification task; multi-class $m = 1$ and multi-label $m = |L|$, where $|L|$ indicates the number of labels of the label space.

probabilities (Equation B.1). We vary the components (ii) and (iv) to enforce better comparability between the models; component (i) depends on the selection for (ii), and (v) depends on the type of classification task. For the multi-class experiments, we utilize the softmax activation function and select the class with the maximum probability as the prediction $\hat{y}$ for a given document. In contrast, for the multi-label classification experiments, we utilize the sigmoid activation function to perform $|L|$ binary classifications to determine whether the l^{th} medical code $y_i \in \{0,1\}^l$ should be assigned to a given discharge summary.

$$\hat{y}(C, W, B) = act(C \cdot W^\top + B) \tag{3.1}$$

3.3.3 Deformable Phrase-Level Attention Mechanism

As a novel attention mechanism, DPLA learns the final context vector representation for a document by leveraging word-level lexical and phrase-level contextual information to capture the full linguistic meaning of a word [42]. In contrast, traditional attention mechanisms neglect context when learning global relationships between individual words missing important local semantic information [85]. Therefore, we hypothesize that a module that can accurately extract lexical and contextual information will lead to richer context vector representations for a document and to improved model performance. To capture both the lexical and contextual meaning of a word, the DPLA either processes the latent document representation H directly or passes H to a phrase-level extraction module that dynamically estimates the vector representation of the context for each word in a text sequence. The architecture of the DPLA consists of a context range module adopted from work by Ma et al. [92], a mask generation, a masked self-attention mechanism [142], and two attention mechanisms designed to identify the important parts of a document at both the word and phrase levels (Figure 3.2).

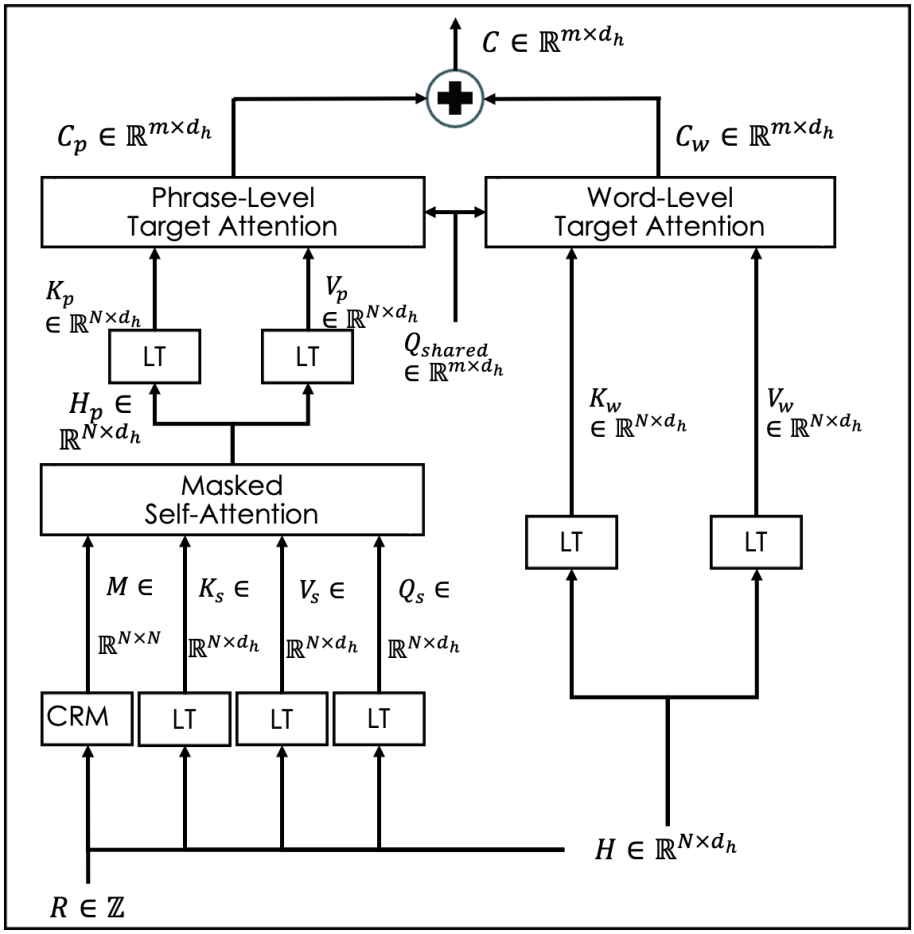


Figure 3.2: Architecture and flow of the deformable phrase-level attention mechanism. The latent document representation H is the output of any given text-encoder architecture and functions as the input to the linear transformations (LT) that generate keys K, values V, and queries Q and as input to the context range module (CRM). The CRM uses H and a predetermined maximum context range constant R to determine the word-specific context sizes stored in the attention mask M. The masked self-attention mechanism utilizes M, K_s, V_s, and Q_s to estimate the word-specific vector representations at the phrase-level H_p. Both H and H_p are linearly transformed before being fed into two independent target- or label-attention mechanisms with a shared query matrix with m elements to determine the important segments in a clinical document at the word and phrase levels, thereby generating C_w and C_p, respectively. The final document context vector representation C is the element-wise summation of C_w and C_p.

Context Range Module and Attention Mask Generation

The context range module utilizes H as the input for two linear feed-forward layers (Equation 3.2) with parameters $W_1 \in \mathbb{R}^{d_h \times d_h}$, $W_2 \in \mathbb{R}^{d_h \times 1}$, and $b_1, b_2 \in \mathbb{R}^{d_h}$ to dynamically generate $g \in \mathbb{R}^N$, which stores the variable context size for each word in the sequence. Each element in g represents the context size for the i^{th} word in the text sequence with a maximum context range of R. R is a predetermined integer variable that defines the maximum size of a given context window centered around a word to $2R + 1$.

$$g = sigmoid((H \cdot W_1^\top + b_1) \cdot W_2^\top + b_2) \cdot R \tag{3.2}$$

We utilized g as the input to determine the context boundaries for each center word in the attention mask $M(g) \in \mathbb{R}^{N \times N}$; note that the context windows are symmetric except at both tails of a text sequence. The value of each element in M is defined by the distance between the j^{th} word in the sequence to the i^{th} center word (Equation 3.3).

$$M(g)_{i,j} = \begin{cases} 0 & |i-j| \leq g_i, \qquad \forall i, j \in [0, N], \\ -\infty & else \end{cases} \tag{3.3}$$

Phrase-Level Context Estimation using Masked Self-Attention

The DPLA uses a masked self-attention mechanism [142] to accurately estimate the vector representation for each word-specific context. The module first transforms the latent document representation H into the keys $K_s \in \mathbb{R}^{N \times d_h}$ (Equation 3.4), values $V_s \in \mathbb{R}^{N \times d_h}$ (Equation 3.5), and queries $Q_s \in \mathbb{R}^{N \times d_h}$ (Equation 3.6) with weights $W_K, W_V, W_Q \in \mathbb{R}^{d_h \times d_h}$ and biases $b_K, b_V, b_Q \in \mathbb{R}^{d_h}$. The three matrices K_s, V_s, and Q_s and the attention mask M are then fed into the masked self-attention mechanism (Equation 3.7) to generate the latent document representation $H_p \in \mathbb{R}^{N \times d_h}$. Each element in H_p represents the context vector representation of the i^{th} word as the

weighted sum of all the word-embedding vectors inside a given context window. The purpose of this masked attention mechanism is to ensure that only neighboring words within the defined context window contribute to the self-attention operations for each center word, thereby enforcing phrase-level interpretation for each word.

$$K_s = ELU(H \cdot W_K^\top + b_K) \tag{3.4}$$

$$V_s = ELU(H \cdot W_V^\top + b_V) \tag{3.5}$$

$$Q_s = ELU(H \cdot W_Q^\top + b_Q) \tag{3.6}$$

$$H_p(Q_s, K_s, V_s, M) = softmax(\frac{Q_s \cdot K_s^\top}{\sqrt{d_h}} + M)V_s \tag{3.7}$$

Word- and Phrase-Level Attention and Final Document Context Vector Generation

The latent document vector representations H and H_p are passed into a final attention mechanism—target attention [50] for multi-class or label attention [99] for multi-label models—to determine the most relevant parts of a document at the word and phrase levels (Equation 3.8). This attention mechanism first transforms the word-level H and phrase-level H_p latent document representations into the key-value pair matrices $K_w, V_w \in \mathbb{R}^{N \times d_h}$ and $K_p, V_p \in \mathbb{R}^{N \times d_h}$ by utilizing Equations 3.4 and 3.5 and their respective parameters. Both key matrices are then compared with a shared query matrix $Q_{shared} \in \mathbb{R}^{m \times d_h}$ that contains either task-specific ($m = 1$) reference information in the multi-class or label-specific ($m = |L|$) reference information in the multi-label classification setting. The following output is then aligned with a softmax function and multiplied with the remaining value matrices to generate the final word- and phrase-level context vector representations $C_w, C_p \in \mathbb{R}^{m \times d_h}$. Last, we perform an element-wise summation of C_w and C_p to generate the final document context vector representation C.

$$C_i(Q_{shared}, K_i, V_i) = softmax(\frac{Q_{shared} \cdot K_i^\top}{\sqrt{d_h}})V_i \qquad i \in w, p \tag{3.8}$$

3.3.4 Experimental Design

Multi-Class Experiments on Electronic Pathology Reports

Population-level cancer surveillance is a critical tool for combating cancer. However, a significant hindrance to effectively leveraging this surveillance is the long time gap (i.e., 24–48 months [130]) between the first diagnosis of cancer in a patient and its inclusion in population-level databases. Caused by manual processes in the surveillance framework, this time gap must be shortened and ideally moved closer to (near) real-time curation of cancer data. To accomplish this goal, we developed a model to extract critical cancer elements from electronic pathology reports.

Proposed Model Architecture: The proposed model architecture consists of a transformer-based text-encoder architecture and our DPLA, denoted as CLF-DPLA. We selected the CLF that was pretrained on general medical text as the text-encoder architecture because it was effective in classifying clinical documents and outperformed other contemporary transformers and conventional text-encoder architecture-based models (e.g., CNNs, Bi-GRUs) on a variety of clinical text classification tasks [84, 95]. Instead of following standard practice with transformers and using the hidden states of the first special token $<s>$ or $[CLS]$ of the encoded sequence as the input to the classification layer, we pass the hidden states of the entire sequence to our proposed DPLA to generate a final document context vector representation.

Baseline Models: We consider three baseline models for the evaluation of our proposed CLF-DPLA: (i) a shallow CNN adapted to multitask learning (MT-CNN)

[74, 7]; (ii) the hierarchical self-attention network adapted to multitask learning (MT-HiSAN) [50], which is the main model currently used in practice by the NCI registries for information extraction tasks for cancer pathology reports; and (iii) the base version of the CLF, which uses the vector representation of the first special token of the text sequence as the final output vector representation (CLF-BS) [84]. For more information on the baseline models, see our supplemental material in Appendix A.1.

Multi-Label Experiments on Hospital Discharge Summary Notes

The ultimate goal of public health agencies is the near real–time, population-level health surveillance and health security [**?** 1] across all diseases. Therefore, MIMIC-III [69] is an ideal application because it represents real-world patient cases that require the information extraction of multiple medical concepts per document to evaluate AI-driven clinical text classification models.

Text-Encoder Architectures Evaluated with the DPLA: We evaluated the performance of three commonly used text-encoder architectures with our proposed DPLA [95, 89]: (i) a shallow CNN [74], (ii) a two-level Bi-GRU [31], and (iii) the transformer-based CLF [84] using the hidden states of the entire sequence as the input for the DPLA (see Appendix A.1).

Baseline Models: We utilized the same three text-encoder architectures—CNN [74], Bi-GRU [31], and CLF [84] (see Appendix A.1)—with the label-wise attention mechanism [99] instead of the DPLA as the main baseline models and provided scores for previous and current state-of-the-art models. The label-wise attention mechanism has been used consistently as a core component in state-of-the-art models for this clinical text classification task [87, 99, 143, 88].

Quantitative Evaluation

We evaluated the performance of all models by using quantitative metrics established in the clinical text classification literature and measured the micro-[†] and macro-averaged F1-score (Equation B.3) [49, 76, 97]. For the multi-label experiments, we also provided the micro- and macro-averaged AUC (area under the curve) scores and precision at k for k in$\{5, 8, 15\}$. We used the widely accepted normal approximation [115] to compute 95% confidence intervals over five runs (see Appendix A.1).

Trustworthiness: Evaluation of Model Robustness and Predictive Confidence

To gain the trust of clinicians, AI-driven clinical support systems must output their confidence levels when making decisions and be *accurate* when they are highly confident. Furthermore, ideally, the high confidence levels of such models should also translate to out-of-distribution data—data originating from sources not originally included in the training regimen. Therefore, we retrospectively evaluated the output of both the ID and OOD experiments on the pathology reports by applying soft-abstention [35]. Soft-abstention determines a probability-based rejection rule to retrieve the retention proportion of non-abstained documents with confident and correct predictions while maintaining a low error rate; this error rate was set to 3% to approximate human error rates. See Appendix A.2 for more details on the soft-abstention procedure.

Model Training and Hyperparameter Optimization

We trained all models on high-performance computing clusters using automatic mixed-precision and data parallelism frameworks to accelerate training. We optimized

[†]We note that in multi-class classification tasks the micro-averaged F1-score equals the performance accuracy metric. Therefore, we will refer to performance accuracy when describing the performance results on the SEER dataset.

the hyperparameters of all evaluated models on each dataset by using the hill-climbing strategy (see Appendix A.1).

3.4 Results

3.4.1 Multi-Class Experiments on SEER Electronic Pathology Reports

In-Distribution and Out-of-Distribution Experiments

Table 3.1 summarizes the performance accuracies and macro-averaged F1-scores with 95% confidence intervals of all models evaluated on the ID and OOD electronic pathology reports. The significant findings of the ID experiments are as follows: (i) both CLF-based models significantly outperformed the MT-HiSAN and the slightly worse MT-CNN, and (ii) extending the CLF with our proposed DPLA further enhanced the performance of the transformer model across all five tasks. Specifically, the CLF-DPLA significantly outperformed all three baseline models across all five tasks, except for behavior, for which the improvement in accuracy was only marginal. Specifically, the CLF-DPLA improved the performance accuracy and F1-macro scores, respectively, across all tasks on average by 0.10 and 0.025 for the CLF-BS, 0.035 and 0.048 for the MT-HiSAN, and 0.039 and 0.083 over the MT-CNN. Compared with their ID counterparts, all models experienced a performance drop in four of the five oncological tasks, including site, subsite, laterality, and histology, but exhibited better performance for behavior on the OOD experiments. Last, our proposed CLF-DPLA outperformed all three baselines across all tasks with an average improvement in performance accuracy and macro-averaged F1-score, respectively, of 0.005 and 0.020 over the CLF-BS, 0.032 and 0.059 over the MT-HiSAN, and 0.035 and 0.077 over the MT-CNN.

Table 3.1: Experimental results for the Multi-task Convolutional Neural Network (MT-CNN), the Multi-task Hierarchical Self-Attention Network (MT-HiSAN), the base Clinical Longformer (CLF-BS), and the CLF extended with the deformable phrase-level attention mechanism (CLF-DPLA) while performing information extraction of cancer data elements from in-distribution and out-of-distribution test electronic pathology reports. Performance accuracy and macro-averaged F1-score are displayed with 95% confidence intervals. Bold scores indicate highest performance and statistically significant difference (*). Out-of-distribution performance increases are indicated by ↑, decreases are indicated by ↓, and no change is indicated by – when compared with the respective in-distribution test results.

Task	Metric	Models			
		MT-CNN	MT-HiSAN	CLF-BS	CLF-DPLA
In-Distribution Data (n = 78,211)					
Site	Accuracy	0.918 (0.916, 0.920)	0.923 (0.922, 0.925)	0.941 (0.939, 0.942)	**0.946*** **(0.944, 0.947)**
	F1-Macro	0.664 (0.660, 0.667)	0.697 (0.694, 0.700)	0.711 (0.708, 0.715)	**0.734*** **(0.731, 0.737)**
Subsite	Accuracy	0.657 (0.653, 0.660)	0.665 (0.662, 0.668)	0.703 (0.700, 0.707)	**0.720*** **(0.717, 0.724)**
	F1-Macro	0.317 (0.314, 0.320)	0.356 (0.353, 0.360)	0.368 (0.364, 0.371)	**0.388*** **(0.385, 0.392)**
Laterality	Accuracy	0.897 (0.895, 0.900)	0.901 (0.899, 0.903)	0.927 (0.925, 0.929)	**0.939*** **(0.937, 0.940)**
	F1-Macro	0.501 (0.498, 0.505)	0.545 (0.542, 0.548)	0.574 (0.570, 0.577)	**0.590*** **(0.587, 0.594)**
Histology	Accuracy	0.759 (0.756, 0.762)	0.762 (0.758, 0.765)	0.797 (0.794, 0.800)	**0.811*** **(0.808, 0.813)**
	F1-Macro	0.269 (0.266, 0.272)	0.328 (0.325, 0.332)	0.372 (0.369, 0.376)	**0.411*** **(0.408, 0.414)**
Behavior	Accuracy	0.963 (0.961, 0.964)	0.962 (0.961, 0.964)	0.973 (0.972, 0.975)	**0.976** **(0.975, 0.977)**
	F1-Macro	0.850 (0.847, 0.852)	0.847 (0.844, 0.850)	0.863 (0.861, 0.865)	**0.890*** **(0.888, 0.892)**
Out-of-Distribution Data (n = 106,497)					
Site	Accuracy	0.915 ↓ (0.913, 0.917)	0.920 ↓ (0.919, 0.922)	0.938 ↓ (0.937, 0.940)	**0.942*** ↓ **(0.941, 0.944)**
	F1-Macro	0.648 ↓ (0.646, 0.651)	0.673 ↓ (0.670, 0.676)	0.684 ↓ (0.681, 0.687)	**0.704*** ↓ **(0.701, 0.707)**
Subsite	Accuracy	0.651 ↓ (0.648, 0.654)	0.652 ↓ (0.650, 0.655)	0.701 ↓ (0.699, 0.704)	**0.713*** ↓ **(0.710, 0.716)**
	F1-Macro	0.307 ↓ (0.305, 0.310)	0.290 ↓ (0.288, 0.293)	0.368 – (0.365, 0.370)	**0.388*** – **(0.385, 0.391)**
Laterality	Accuracy	0.900 ↑ (0.898, 0.902)	0.903 ↑ (0.902, 0.905)	0.928 ↑ (0.927, 0.930)	**0.932*** ↓ **(0.931, 0.934)**
	F1-Macro	0.493 ↓ (0.490, 0.496)	0.521 ↓ (0.518, 0.524)	0.545 ↓ (0.542, 0.548)	**0.558*** ↓ **(0.555, 0.561)**
Histology	Accuracy	0.754 ↓ (0.755, 0.760)	0.757 ↓ (0.755, 0.760)	0.793 ↓ (0.791, 0.796)	**0.796** ↓ **(0.794, 0.799)**
	F1-Macro	0.262 ↓ (0.259, 0.264)	0.312 ↓ (0.309, 0.315)	0.364 ↓ (0.361, 0.367)	**0.397*** ↓ **(0.394, 0.400)**
Behavior	Accuracy	0.967 ↑ (0.966, 0.968)	0.969 ↑ (0.968, 0.970)	0.978 ↑ (0.977, 0.978)	**0.979** ↑ **(0.978, 0.979)**
	F1-Macro	0.847 ↓ (0.845, 0.850)	0.849 ↑ (0.847, 0.851)	0.883 ↑ (0.881, 0.885)	**0.895*** ↑ **(0.893, 0.897)**

In-distribution data collected from NCI SEER registries in Kentucky, Louisiana, New Jersey Washington and Utah. Out-of-distribution collected from NCI SEER registry in New Mexico. The 95% confidence interval is computed via normal approximation [115].

Table 3.2: Retention proportion (RP) results of soft-abstention with an error rate of 3% on in-distribution and out-of-distribution before unseen cancer pathology reports.

Task	Metric	Models			
		MT-CNN	MT-HiSAN	CLF-BS	CLF-DPLA
In-Distribution Data ($n = 78,211$)					
Site	RP in %	90.0	91.3	94.2	95.3
	Threshold	0.686	0.639	0.694	0.669
	Accuracy	0.970	0.970	0.970	0.971
Subsite	RP in %	34.8	37.3	40.0	41.6
	Threshold	0.900	0.872	0.941	0.952
	Accuracy	0.967	0.966	0.967	0.970
Laterality	RP in %	82.5	84.6	90.8	93.5
	Threshold	0.806	0.736	0.727	0.695
	Accuracy	0.969	0.968	0.968	0.968
Histology	RP in %	26.3	28.7	32.2	34.9
	Threshold	0.957	0.924	0.967	0.965
	Accuracy	0.971	0.971	0.971	0.971
Behavior	RP in %	98.2	98.4	100	100
	Threshold	0.609	0.591	0	0
	Accuracy	0.970	0.969	0.973	0.976
Out-of-Distribution Data ($n = 106,497$)					
Site	RP in %	89.4↓	91.0↓	93.8↓	94.8↓
	Threshold	0.688	0.634	0.694	0.669
	Accuracy	0.969	0.967	0.969	0.969
Subsite	RP in %	31.8↓	34.6↓	39.9↓	41.9↑
	Threshold	0.901	0.870	0.941	0.952
	Accuracy	0.963	0.960	0.958	0.956
Laterality	RP in %	82.3↓	85.4↑	91.0↑	92.4↓
	Threshold	0.806	0.736	0.727	0.695
	Accuracy	0.969	0.968	0.970	0.969
Histology	RP in %	17.2↓	24.7↓	28.8↓	32.2↓
	Threshold	0.957	0.925	0.967	0.965
	Accuracy	0.969	0.964	0.966	0.964
Behavior	RP in %	98.2 –	98.5↑	100 –	100 –
	Threshold	0.630	0.591	0	0
	Accuracy	0.975	0.975	0.978	0.979

Soft-Abstention: Predictive Confidence

Table 3.2 summarizes the results of the soft-abstention analysis across all models and oncological tasks. The main findings of the soft-abstention analysis for the ID and OOD test datasets are as follows: (i) the retention proportions (RP) for all models depend on the complexity of a given oncological task and show large RP (>80%) for site, behavior, and laterality and low RP (<42%) for the more complex tasks of subsite and histology; (ii) the majority of RP decreased on the OOD dataset compared with their ID-based counterparts; (iii) utilizing the transformer-based CLF as the text-encoder architecture provided the strongest initial boost in RP; (iv) our proposed CLF-DPLA achieved the largest RP across all tasks and test datasets; and (v) the CLF-DPLA substantially increased the RP on the subsite and histology tasks over the previous state-of-the-art MT-HiSAN for both the ID (4% and 6%) and OOD (7% and 8%) results.

The complex phenotyping of cancer is considered a high-risk scenario that requires the simultaneous and confident classification of multiple cancer characteristics from a clinical text document, i.e., a single classification would not be sufficient. Therefore, we computed the intersection of documents with complete phenotyping and report the RP (i.e., documents with predictions that meet the probability-based rejection thresholds across all oncological tasks). Because various clinical applications have different error rates, we simulate the proportion of ID and OOD reports with complete phenotyping for human error rates ranging from 1% to 10% (Figure 3.3).

Generally, the RP increases rather linearly with rising error rates across all models. The RP difference between the models becomes more pronounced with rising error rates. Beginning with an error rate of 0.04 across both datasets, the CLF-based models achieve larger retention proportions than both conventional models, thereby indicating that the CLF-based models provide predictions with larger probabilities for a larger subset of documents than the MT-HiSAN and MT-CNN. The results further indicate that the DPLA helps the CLF to retain more documents. With the

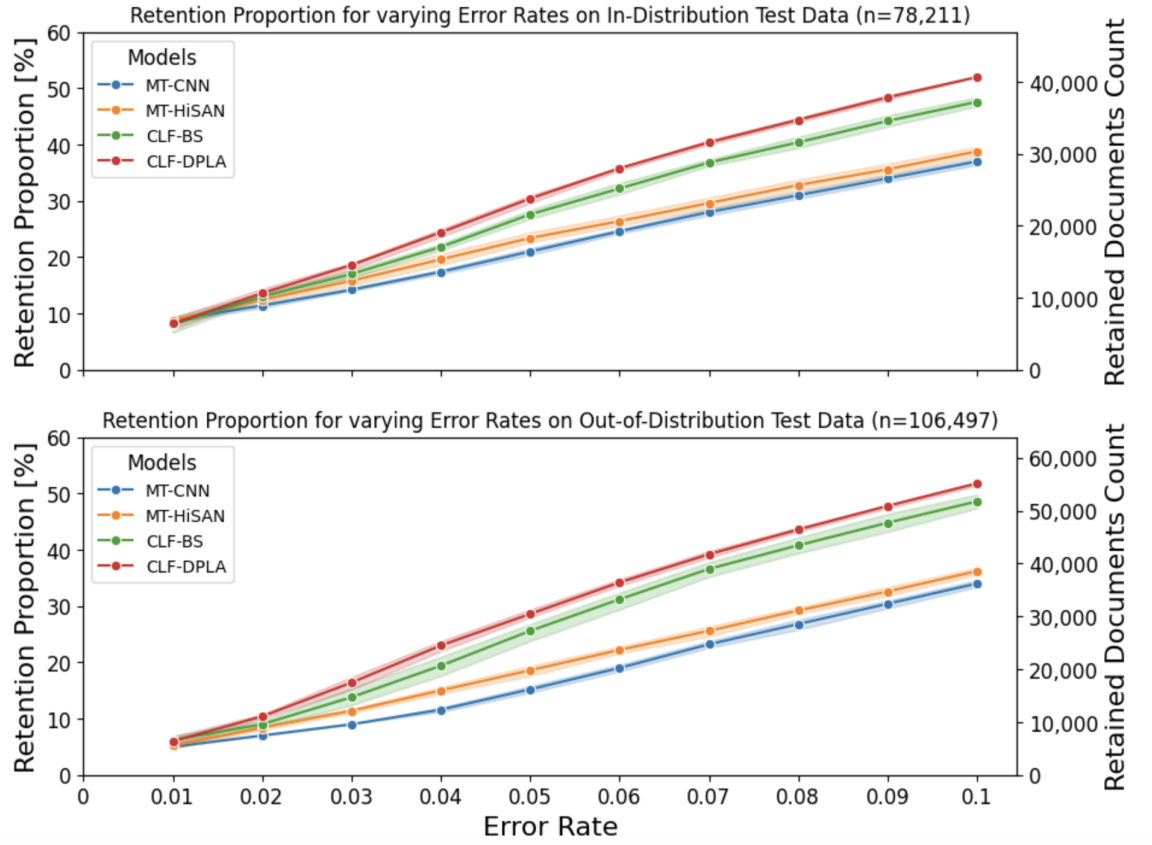


Figure 3.3: Sensitivity analysis of the retention proportion for different error rates on the in-distribution and out-of-distribution test data comprising SEER-based electronic pathology reports. The results depict the varying degrees of retention proportion for four models, including the multitask convolutional neural network (MT-CNN), multitask hierarchical self-attention network (MT-HiSAN), the baseline Clinical Longformer (CLF-BS), and the Clinical Longformer extended with the deformable phrase-level attention mechanism (CLF-DPLA).

exemption for 0.01, the CLF-DPLA achieved the largest RP across all error rates and both test datasets.

3.4.2 Results: Multi-Label Experiments on Hospital Discharge Summary Notes

The MIMIC-III-Full and MIMIC-III-50 results for the text-encoder architectures CNN, Bi-GRU, and transformer-based CLF extended with either the label-wise attention mechanism or our DPLA are shown in Tables 3.3 and 3.4. Generally, all three text-encoder architectures extended with the DPLA achieve similar or better performance than their counterparts that uses label-attention. For MIMIC-III-Full, the CNN benefited from using the DPLA by achieving minor performance improvements on the micro-averaged F1-score and significant performance improvements on the macro-averaged F1-score. In contrast, both contextual-based text-encoder architectures (Bi-GRU and CLF) achieved better performance with the label-wise attention mechanism than with our DPLA. For MIMIC-III-50, all three text-encoder architectures achieved marginal or significant performance improvements with the DPLA over their respective baselines. For all three text-encoders, the performance gains are more significant on the macro-averaged F1-scores than on the micro-averaged F1-scores. In contrast to the MIMIC-III-Full results, the strongest improvements on the MIMIC-III-50 were achieved on the Bi-GRU and the CLF text-encoders.

Table 3.3: Experimental results on MIMIC-III-Full for the convolutional neural network (CNN), the bi-directional gated recurrent unit neural network (Bi-GRU), and the transformer-based Clinical Longformer (CLF) utilizing label-attention or the deformable phrase-level attention mechanism (DPLA); the label-attention results were taken from previous work [95]. Bold indicates the better performing attention mechanism for each model.

Model	Attention Mechanism	AUC		F1		P@k	
		Macro	Micro	Macro	Micro	8	15
CNN	Label	0.862 (0.850, 0.873)	0.979 (0.975, 0.984)	0.041 (0.034, 0.048)	0.468 (0.451, 0.485)	0.645 (0.629, 0.662)	0.494 (0.477, 0.510)
	DPLA	**0.876 (0.865, 0.887)**	**0.981 (0.977, 0.986)**	**0.059 (0.051, 0.067)**	**0.492 (0.476, 0.509)**	**0.668 (0.652, 0.684)**	**0.519 (0.502, 0.536)**
Bi-GRU	Label	**0.890 (0.880, 0.901)**	**0.984 (0.980, 0.989)**	**0.066 (0.058, 0.075)**	**0.531 (0.514, 0.548)**	0.702 (0.687, 0.717)	0.548 (0.531, 0.565)
	DPLA	0.886 (0.875, 0.896)	**0.984 (0.980, 0.988)**	0.061 (0.053, 0.070)	0.523 (0.506, 0.540)	**0.706 (0.691, 0.722)**	**0.552 (0.535, 0.569)**
CLF	Label	**0.903 (0.893, 0.913)**	**0.985 (0.981, 0.989)**	0.057 (0.049, 0.065)	**0.524 (0.507, 0.541)**	**0.724 (0.709, 0.739)**	**0.571 (0.554, 0.588)**
	DPLA	0.888 (0.877, 0.898)	0.982 (0.977, 0.986)	**0.061 (0.053, 0.069)**	0.488 (0.471, 0.504)	0.678 (0.662, 0.694)	0.526 (0.510, 0.543)

95% confidence interval computed via normal approximation following [115].

Table 3.4: Experimental results on MIMIC-III-50 for the convolutional neural network (CNN), the bi-directional gated recurrent unit neural network (Bi-GRU), and the transformer-based Clinical Longformer (CLF) utilizing label-attention or the deformable phrase-level attention mechanism (DPLA); the label-attention results were taken from previous work [95]. Bold indicates the better performing attention mechanism for each model.

Model	Attention Mechanism	AUC		F1		P@k
		Macro	Micro	Macro	Micro	5
CNN	Label	**0.893 (0.882, 0.922)**	**0.922 (0.913, 0.931)**	0.556 (0.539, 0.573)	0.632 (0.616, 0.648)	**0.628 (0.611, 0.644)**
	DPLA	0.890 (0.880, 0.901)	0.917 (0.908, 0.927)	**0.587 (0.571, 0.604)**	**0.648 (0.632, 0.664)**	0.623 (0.607, 0.640)
Bi-GRU	Label	0.883 (0.872, 0.894)	0.915 (0.905, 0.924)	0.529 (0.512, 0.545)	0.624 (0.608, 0.640)	0.616 (0.600, 0.633)
	DPLA	**0.895 (0.885, 0.905)**	**0.923 (0.914, 0.932)**	**0.577 (0.561, 0.594)**	**0.646 (0.630, 0.662)**	**0.633 (0.617, 0.650)**
CLF	Label	0.901 (0.891, 0.911)	0.925 (0.917, 0.934)	0.590 (0.573, 0.606)	0.651 (0.635, 0.667)	0.640 (0.624, 0.657)
	DPLA	**0.910 (0.900, 0.919)**	**0.931 (0.922, 0.939)**	**0.634*** **(0.618, 0.650)**	**0.682 (0.666, 0.698)**	**0.650 (0.634, 0.666)**

95% confidence interval computed via normal approximation following [115].

3.5 Discussion

3.5.1 Performance Improvements through Context Estimation

We evaluated our proposed DLPA mechanism on two clinical text classification tasks that involved the extraction of critical cancer data elements from SEER electronic pathology reports and the automated medical encoding of MIMIC-III hospital discharge summary notes. Our experiments on both classification tasks showed that extending either conventional or transformer-based text classification models with our DPLA can yield significant performance improvements.

This finding suggests that enabling an attention mechanism to simultaneously learn lexical and contextual information for a word can lead to improved model performance over attention mechanisms that consider only word-level information. Our experimental results on the MIMIC-III showed that the all three evaluated text-encoder architectures (CNN, Bi-GRU, and transformer-based CLF), when extended with the DPLA, achieve similar or better performance compared with the word-level, label-wise attention mechanism. The DPLA complements the CNN's method of encoding the meaning of a word particularly well, as our DPLA enables the CNN to expand its view of only short-range convolution windows (i.e., multiple n-grams with three, four, and five words) with an additionally larger local context window to provide more semantic information for each word to the model. Nevertheless, the Bi-GRU and CLF showed great results on MIMIC-III-50 but worse performance on MIMIC-III-Full, indicating that the DPLA causes confusion in model training when faced with a large label set. A common problem noted [39] with MIMIC-III-Full is the presence of labels with only a few samples, exacerbating the training of the DPLA due to its more complex architecture.

Despite the performance improvements of the models with DPLA over the models with label-attention on MIMIC-III, none could match the performance accuracy of

current state-of-the-art models (see Appendix A.2). Although these state-of-the-art models incorporate label-attention, we think the main difference in performance stems from their use of more complex text-encoder architectures versus our tested architectures that enable the models to learn more meaningful word-level encodings for the sequence. This suggests that the quality in word-level encodings impacts an attention mechanism's ability to find important parts in a sequence. Given our results on the usage of either DPLA or label-attention, we think that models could further improve their performance by using the DPLA, which is flexible, instead of a label-attention mechanism.

One aspect of the DPLA is the use of dynamically generated variable (V)-sized context windows instead of fixed (F)-sized context windows in the attention mask to account for contextual differences between words. On 25 non-medical datasets, Ma et al. [92] showed that deploying V-sized context windows is preferable to deploying F-sized context windows. However, clinical text has different characteristics. Therefore, we performed additional experiments on both clinical text datasets (see Appendix A.2). The results indicate that selecting V- or F-sized context windows depends on the clinical text classification task; note that the performance differences were marginal. However, we think that using V-sized context windows reduces time and resources spent on hyperparameter optimization given that models have shown to dynamically converge to an appropriate context size range for a given oncological information extraction task (e.g., smaller for site and larger for histology).

3.5.2 Task-Dependent Word-Level Context Sizes

In our approach, we estimated the size of a word's context and effectively the size of a phrase by using the DPLA's context range module. We know from the literature that the linguistic meaning of a word is defined by itself and its context and that the context size may vary between words. We hypothesize, however, that a word's context size will vary due to the idiosyncratic characteristics of a given oncologic task.

In particular, we would expect the context size, defined as the words before and after a given word, to be small for site and longer for histology. To test our hypothesis, we analyzed the frequency distributions of the word-level context sizes of the three most important phrases extracted from correctly predicted reports for task site and histology. Figure 3.4 shows the frequency distributions of a word's context size for site and histology, for which the total context size can range from 3 to 21 tokens long.

The results support our hypothesis and show that the context sizes are small for site and large for histology. For site, almost half of the phrases had a context size of five (i.e., two words before and two words after the center word) and roughly a third of the phrases consisted of either three or seven words. In contrast, the phrases extracted from the histology tasks primarily range between 9 and 19 tokens, in which the largest proportion of phrases, one out of five, has a context size of 17. We believe the context size differences between both tasks stem from a varying degree of complexity in the type of information that a model tries to retrieve (i.e., retrieving primary cancer site information is simpler than retrieving histology information).

In particular, histology is more complex because determining the histological type often require additional information from the text (e.g., behavior). Histology can also have a very large label space with classes that frequently share the same linguistic and lexical information (e.g., morphemes). For example, the word *carcinoma* linguistically functions as a free morpheme because it can stand alone like *Carcinoma, NOS* (8010) or is part of the medical terminology of another histological type such as *Adenomacarcinoma, NOS* (8140). Additionally, common cell types may appear in the textual descriptions of multiple histology types such as *Papillary Carcinoma, NOS* (8050). In comparison, the linguistic complexity for the primary cancer site is low because the medical terms are short and rarely share the same lexical information.

Finally, we want to point out that some part of the context size difference between site and histology may be attributed to the byte-pair encoded (BPE) tokenization scheme used to preprocess the reports, an artifact of using a transformer model. For example, the BPE tokenizer splits the word *carcinoma* into four subwords (*c*, *arc*, *in*,

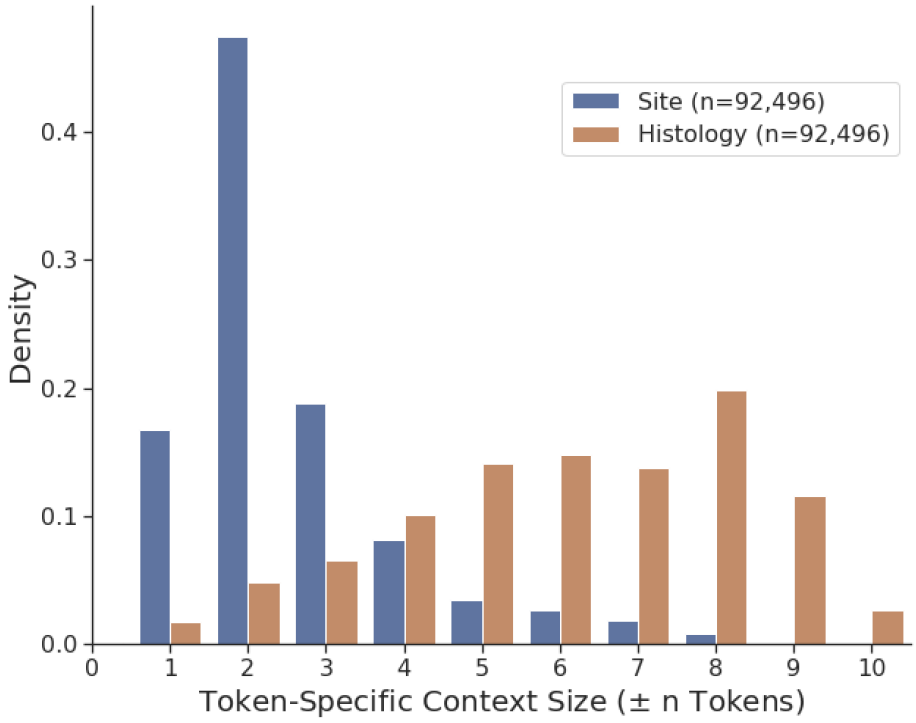


Figure 3.4: Frequency distribution of the token-specific context sizes based on the deformable phrase-level attention mechanism. Depicted here are the context sizes, defined as the number of tokens before and after a given center word, for the three most important phrases that were retrieved from the correctly predicted non-abstained, in-distribution electronic pathology reports ($n = 30,832$) for site and histology.

and *oma*), whereas *breast* or *prostate* are kept intact; note that some primary cancer sites such as *pancreas* may also be broken down into subwords (*p*, *anc*, *re* and *as*).

3.5.3 Model Robustness and Predictive Confidence

The CLF-DPLA was initially designed to improve the automated extraction of cancer information from electronic pathology reports. The potential real-world use of the CLF-DPLA, however, requires robustness against distributional shifts and sufficient predictive confidence in auto-encoding of pathology reports. Electronic pathology reports are collected from multiple cancer registries, and this results in differences in their feature distributions. These differences may degrade the performance of the

CLF-DPLA when trying to classify reports from registries that were not part of the training regimen, thereby reducing the impact on patient outcome during deployment [134]. Unfortunately, frequent retraining of the CLF-DPLA would become financially burdensome and environmentally taxing [137] due to the required computational effort.

Our OOD experiments show that the CLF-DPLA substantially outperformed the MT-HiSAN, which is the current model deployed by NCI registries to extract cancer data from pathology reports, on all oncological tasks. This indicates that the CLF-DPLA may be better at handling shifts in distributional features. Additionally, the soft-abstention analysis showed that the CLF-DPLA retains the largest proportion of pathology reports with complete phenotyping for various simulated error rates. Both results support the notion that our proposed model makes a promising candidate for real-world deployment in the curation of critical cancer elements from electronic pathology reports.

3.5.4 AI-Driven Automated Data Harmonization to Populate Common Data Models

Researchers aiming to foster population-level health improvements by using systematic analyses require an accessible and interoperable network of harmonized data sources. Data harmonization attempts to bring heterogeneous data into consistent formats known as CDMs [46]. Although there are significant efforts to populate CDM frameworks (e.g., the Observational Medical Outcomes Partnership [58], mCODE [106]), we think AI-based models can automate the harmonization and curation of CDMs, thereby increasing data quality and reducing processing time. In particular, we think that our proposed CLF-DPLA method showed favorable properties for the automatic extraction of critical cancer data elements—especially given that the majority of this information is still stored in unstructured text [153]—from pathology reports collected from different cancer registries and can consistently

and rapidly populate primary and secondary cancer condition fields in the mCODE CDM, thereby enabling oncologists to perform time-sensitive personalized medicine.

3.5.5 Limitations

Finally, we want to note this study's limitations. One limitation is the use of pathology reports from a single registry for the OOD dataset, which does not truly reflect the potential real-world patient demographics. Another limitation is that we did not discuss the DPLA's potential to function as a source of explainability. We also want to note that the current literature describes the use of attention scores for explainability of a model's predictions as critical [16]. This is because the attention scores do not correlate well with gradient-based important measures [66] or their ability to deceive humans annotators [112].

3.6 Conclusion

This work proposes a novel deformable phrase-level attention mechanism that extracts key information at the word and phrase levels to improve the performance of clinical text classification models. We showcase the achievable performance improvements of our method on two datasets of clinical text. Specifically, our DPLA-enhanced models achieve new state-of-the-art performance on the information extraction of critical cancer data elements from SEER-based electronic pathology reports. We show that the DPLA outperforms common word-level attention mechanisms, is robust against distributional shifts in the input, improves proportion of documents predicted accurately with high confidence, and has potential for automated data harmonization and curation of common data models[‡].

[‡]The code for this project is available at https://github.com/cmetzner93/phrase_level_attention/tree/main/src

Chapter 4

Can human clinical rationales improve the performance and explainability of clinical text classification models?

Disclosure Statement

A version of this chapter is planned to be submitted to an appropriate journal in the clinical informatics space.

Author's contributions are as follows: Christoph Metzner designed and executed the study; Shang Gao, Drahomira Herrmannova, and Heidi Hanson contributed to the experimental design and analysis of the results. CM organized, programmed, and performed all experiments. CM wrote the manuscript. All authors read, reviewed, and approved the final manuscript.

Abstract

Objective: AI-driven clinical text classification is vital for the explainable automated retrieval of population-level health information. This work investigates human-based clinical rationales as additional supervision signal to improve the performance and explainability of transformer-based clinical text classification models performing the automated medical encoding of clinical documents.

Materials and Methods: We analyze 99,125 human-based clinical rationales providing plausible explanations for the primary cancer site diagnosis as additional training samples. We investigated sufficiency to measure rationale quality as a rationale preselection method. We evaluated transformer-based models for the extraction of the primary cancer site from 128,649 electronic pathology reports.

Results: Utilizing clinical rationales as additional training data can lead to improved model performance in high-resource scenarios but inconsistent behavior in low-resource scenarios. Using sufficiency as an automatic metric to preselect rationales leads to inconsistent performance results. Models trained on rationales are outperformed by models trained on additional reports.

Discussion: Clinical rationales do not consistently improve model performance and are outperformed by models trained on additional reports. Therefore, if optimizing for accuracy, annotation efforts should focus on labeling more reports rather than labeling rationales. If optimizing for explainability, training the models on data supplemented with rationales may help models better identify rationale-like features.

Conclusion: This work investigates human-based clinical rationales to improve clinical text classification models that perform automated encoding of clinical documents. Using clinical rationales as additional training data results in lower model performance improvements and slightly better explainability, which is measured as the average token-level rationale coverage compared with training on additional reports.

4.1 Background and Significance

AI-driven clinical support systems are revolutionizing patient care by enhancing a clinician's access to detailed patient information and evidence-based recommendations [40]. These systems rely largely on clinical information extracted from an exponentially growing body of unstructured electronic health records (e.g., electronic pathology reports), necessitating efficient automated extraction through machine learning (ML)-based text classification models. To gain the trust of patients and clinicians in high-stakes scenarios such as patient care, these ML models must demonstrate both high performance and explainability. In addition to ongoing algorithmic advances, recent work in non-clinical domains utilizes human-based rationales as an additional supervision signal to improve model performance and explainability [146]. Although initial results are promising, it remains unclear if human-based *clinical* rationales will enable similar model improvements in clinical text classification tasks, including the automated encoding of clinical documents. Clinical documents are particularly challenging to process owing to their use of complex clinical language (e.g., jargon, abbreviations, varying terminology) and

multiple contextual topics (e.g., patient history information, institutional details, diagnostic information), both of which exacerbate the difficulties associated with learning from human-based clinical rationales (hereafter referred to as *rationales*). To this end, this work investigates how clinical text classification models are impacted by incorporating rationales in the model training.

Rationales are concise, plausible explanations that adequately justify a document's ground-truth label, such as a clinical diagnosis [26, 146]. Rationales are either (i) abstractive free-form text created by humans or generative models [28, 157, 78, 121] or (ii) extractive text highlights retrieved by a human [158, 161, 15, 136, 163, 9, 160] or a model [125, 140, 162, 111]. Although some research has explored the use of rationales to enhance model explainability [136, 9, 162, 111, 54, 116], the primary application of rationales lies in improving the performance of ML models [158, 161, 15, 125, 9, 140, 162, 116]. Generally, researchers leverage rationales as a form of prior domain knowledge in three key ways: (i) as supplementary samples that enable models to discover new connections between input features and output labels [158, 125, 116], (ii) as fine-grained supervision to provide additional ground-truth labels at the word or sentence level [161, 9, 140], and (iii) as guidance for internal model components (e.g., attention mechanisms) to direct a model's focus toward rationale-like features [15, 163, 9, 160, 111]. However, more recent work has explored large language models (LLMs) for generating abstractive rationales to help the model to self-rationalize its output [28, 157, 78, 121].

While previous work on rationales in the clinical domain explored generative approaches [140, 78, 121], our study focuses on extractive rationales to enhance specialized discriminative models for the automated encoding of electronic pathology reports with primary cancer site diagnoses. Interestingly, preliminary experiments showed that rationales used as additional token-level supervision, either as ground-truth labels for token-sequence classification models or as attention masks, resulted in lower performance compared to utilizing them as supplementary independent training samples (Table B.4). These findings, coupled with previous research demonstrating

the effectiveness of rationales as supplementary training input [158, 125] or as the sole training input [116], motivated us to explore the potential of our rationales as supplementary training data. The main idea of this approach is that rationales represent information pruned of noise, thereby providing a direct link to a given primary cancer site [64] and allowing the model to learn more robust relationships between rationale-like input and the label space and thus improve explainability and performance.

4.2 Objective

This study examines how incorporating rationales for primary cancer site diagnoses as an additional supervision signal during model training can enhance transformer-based models for automated clinical document classification. We perform multiple experiments focused on the classification of electronic pathology reports with the primary cancer site diagnosis collected by the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) program from 2011 to 2022 [2]. The experiments gauge the achievable model performance when trained on data supplemented with available rationales in high- and low-resource scenarios—high-resource scenarios assume access to rationales for all classes, whereas low-resource scenarios assume access to rationales for only a few classes. Additionally, we evaluated the performance of models trained on a dataset supplemented with sufficient rationales that are filtered by using the *sufficiency* metric [37, 26]. We contrast these performances against a control experiment involving models trained on a dataset supplemented with additional labeled reports which provides us with additional insight on the optimal return on investment of annotation resources. In high-resource scenarios, the results show improved model performance over the baselines through the supplementation of rationales; however, these rationale-supplemented models are outperformed by the control models trained on larger datasets of reports. In low-resource scenarios, rationale-supplementation can even lead to decreased and

inconsistent performance. Our experiments on filtering out low-quality rationales with the sufficiency metric reflect these inconsistencies. These findings suggest that, although rationales may improve performance in certain scenarios, quality-related issues can have a negative impact on model performance. Therefore, we analyze rationale quality based on existing literature of explainable AI and learning with rationales. Finally, we discuss our findings for annotation efforts, quality interpretations, and potential reasons why rationales do not consistently improve our models. The contributions of this work are as follows:

- We are the first to investigate extractive, human-based clinical rationales for improving clinical text classification models.
- Our results show that supplementing the training data with human-based clinical rationales can improve model performance and explainability in high-resource scenarios.
- Our findings demonstrate that models trained solely on human-based clinical rationales are outperformed by those trained on full and complementary information, highlighting the complexity of electronic pathology reports.
- We show that the sufficiency metric is unreliable for improving model performance when used to quantify rationale quality.
- Based on our results, we recommend that annotation efforts focus on labeling new, additional documents to maximize model performance.

4.3 Materials and Methods

4.3.1 Data

SEER Cancer Pathology Reports

We obtained a total of 128,649 electronic pathology reports from the New Jersey, Kentucky, and Utah cancer registries of the NCI's SEER program (Table 4.1). Each report is associated with a distinct Cancer-Tumor-Case (CTC) ID; each CTC may be linked to multiple reports. Following annotation standards of the North American Association of Central Cancer Registries, all reports were manually annotated by certified tumor registrars (CTRs) with ground-truth labels for multiple critical cancer data elements, including primary cancer site, subsite, laterality, histology, and behavior. However, this study solely focuses on the classification of 69 primary cancer sites to explore the performance impact of integrating rationales in the training of clinical text classification models. A subset of the reports has been annotated by CTRs with rationales that explain the selected primary cancer site diagnosis.

Table 4.1: Descriptive statistics of the SEER cancer pathology reports highlighted with human-created rationales. Each report in the training split is associated with at least one rationale. The test split consists of 19,546 non-annotated reports and 2,446 reports annotated with a single rationale.

Split	Reports			Rationales				
	N	Mean Text Lengths (**s**)		N	Mean Text Lengths (**s**)		Coverage in % (**s**)	
		Word	Subword		Word	Subword	Word	Subword
Train	87,252	594.4 (523.3)	1029.5 (897.0)	96,679	13.6 (16.8)	26.9 (29.4)	4.1 (5.5)	4.3 (5.4)
Val	19,385	540.5 (466.2)	946.5 (817.3)	0	–	–	–	–
Test	22,012	548.8 (485.2)	961.0 (848.1)	2,446	12.7 (15.9)	25.7 (28.3)	3.8 (5.1)	4.1 (5.3)
Total	128,649	578.5 (509.2)	1005.3 (877.9)	99,125	13.6 (16.7)	26.9 (29.4)	4.1 (5.5)	4.3 (5.4)

The maximum rationale length was set to 128 word tokens (i.e., $mean + 2 \times std$); longer rationales were removed. Maximum sequence length was set to 4,096 subword tokens.

Annotation of Human-Based Clinical Rationales and Populating of Data Splits

We used 99,125 rationales for the primary cancer site created by CTRs during quality control of 89,718 manually reviewed, auto-encoded SEER electronic pathology reports. Notably, a single report may have multiple rationales (Table 4.1). During model training, we incorporated each rationale as an independent sample to supplement the training dataset, similar to previous approaches [158, 125, 116]; note that we also trained models on only rationales for a specific analysis. The models had no access to rationales during inference.

We included nearly all reports with at least one rationale in the training split, reserving a small fraction for the testing split so that we could perform model explainability analyses. To populate the validation and testing splits, reports were randomly sampled from the available total population of SEER electronic pathology reports; these reports do not possess any rationales. We ensured that reports associated with a distinct CTC were confined to a single split and that class distributions were consistent across all splits. Lastly, to provide a control to our rationale-trained models, we randomly selected an additional set of pathology reports ensuring no overlap in CTCs and matching the same class distribution as the set of rationales. This control training dataset enables us to investigate whether annotations efforts should be spent on retrieving rationales or labeling new, unlabeled reports. Additional details on the rationales and their generation are available in Appendix B.1 and [138].

4.3.2 Model Architectures

The transformer-based clinical longformer (CLF) was selected as the base text-encoder architecture for all evaluated models. The CLF is pretrained on medical text and has shown favorable performance on clinical text classification tasks [84, 95]. We focused on two primary models: the baseline version of the CLF (CLF-BS) and

the CLF extended with the recently published deformable phrase-level attention mechanism (CLF-DPLA) [94]. We framed the task as a multiclass classification problem in which a model parameterized by θ aims to map a given input text sequence x—report, rationale, or complement—to its corresponding class label y (i.e., primary cancer site). A detailed description of the model architectures and the classification process can be found in Appendix B.1.

4.3.3 Model Evaluation

We evaluated the performance of all models by using quantitative metrics established in the multiclass text classification literature and measured the performance accuracy and the macro-averaged F1-score [76, 49, 97]. We used the widely accepted normal approximation [115] to compute 95% confidence intervals (see Appendix B.2).

4.3.4 Evaluation of Rationales

Rationales can be useful for enhancing model performance and explainability [136]. Explainability of ML models is generally decomposed into two components: (i) faithfulness (i.e., how well the explanation describes the inner mechanisms of a model for a given prediction) and (ii) plausibility (i.e., how well a human understands the provided explanation). Because our rationales are generated by subject matter experts, we assume that they are all highly plausible explanations for a given primary cancer site diagnosis. However, despite their plausibility, rationales may not be faithful. Therefore, it is important to evaluate rationale faithfulness. Previous work evaluated faithfulness by using two automatic metrics, *sufficiency* and *comprehensiveness* [26, 156, 37], which measure a rationale's ability to reproduce the same prediction as the full report and the degree to which a rationale contains all the information relevant to the prediction, respectively. We used the implementations of both metrics proposed by [26].

Sufficiency (Equation 4.1) measures how a rationale enables the model to reproduce the same prediction as the full report by comparing the prediction probabilities of the full report versus only the rationale:

$$\text{Suff}(x, \hat{y}, \alpha) = 1 - max(0, p(\hat{y}|x) - p(\hat{y}|x, \alpha)), \tag{4.1}$$

where $\hat{y} = argmax[p(y|x)]$ and α is a binary mask indicating whether a word belongs to a rationale. In contrast, comprehensiveness (Equation 4.2) measures how well a rationale encapsulates the information relevant for making the correct prediction by comparing the prediction probabilities between the full report and the complement of the rationale.

$$\text{Comp}(x, \hat{y}, \alpha) = max(0, p(\hat{y}|x) - p(\hat{y}|x, 1 - \alpha)) \tag{4.2}$$

A highly comprehensive rationale should contain more relevant information for the correct prediction than its complement and result in more accurate predictions. The literature argues that a rationale is faithful when it has both high sufficiency and comprehensiveness. Finally, we investigate whether sufficiency is suitable as a metric to discriminate between lower- and higher-quality rationales to improve model training and performance. We experimented with a sufficiency threshold of 0.2, which removes the least sufficient rationales, and a threshold of 0.817 represents the 90% percentile of rationales with false predictions for falsely predicted reports; removing roughly 88% of the rationales that led to false predictions in preliminary experiments.

4.4 Results

4.4.1 Model Performance Across Varying Training Regimens

Table 4.2 summarizes the classification accuracies and macro-averaged F1-scores for the CLF-BS and CLF-DPLA trained on varying training data regimens: (i) reports

only, (ii) reports supplemented with the full set of available rationales, (iii) reports supplemented with a subset of sufficient rationales for the two sufficiency thresholds of 0.2 and 0.817, and (iv) reports supplemented with 96,679 additional labeled SEER electronic pathology reports as a control experiment. The results for both evaluated models show that training on reports supplemented with rationales (regimens ii and iii) can improve performance over training solely on reports (regimen i). However, the control experiments showed that training models on more reports with full information (regimen iv) outperformed the baseline (regimen i) and rationale-based models (regimens ii and iii) across both metrics. The CLF-DPLA architecture consistently outperformed the CLF-BS across all training regimens and metrics.

Both models benefited from supplementing the training dataset (regimen i) with either all rationales (regimen ii) or a set of sufficient rationales (regimen iii), but the CLF-BS did not meaningfully improve in accuracy or macro-averaged F1-score when trained on (regimen ii). However, when trained on sufficient rationales

Table 4.2: Comparison of accuracy and F1-macro scores for varying training regimen for the CLF-BS and the CLF extended with the deformable phrase-level attention mechanism (CLF-DPLA) when performing the classification of the primary cancer site in SEER cancer pathology reports. Both models were trained on only reports ($n_{reports} = 87,252$), reports supplemented with human-based clinical rationales ($n_{rationales} = 96,679$), reports supplemented with sufficient rationales, or reports supplemented with additional reports ($n_{reports,additional} = 96,679$) as a control. The set of additional reports follows the same distribution as the full set of rationales. The * indicates significant performance improvements over reports only. The best performance per model is in bold.

Model	Metric	Number of Added Rationales		Number of Added Sufficient Rationales (>Threshold)		Number of Additional Reports
		0	96,679	94,176 (> 0.2)	80,950 (> 0.817)	96,679
CLF-BS	Accuracy	0.885 (0.881, 0.889)	0.888 (0.883, 0.892)	0.890 (0.886, 0.894)	0.888 (0.884, 0.892)	**0.903*** **(0.899, 0.907)**
	F1-Macro	0.567 (0.561, 0.574)	0.571 (0.564, 0.578)	0.586* (0.580, 0.593)	0.580 (0.570, 0.593)	**0.611*** **(0.604, 0.617)**
CLF-DPLA	Accuracy	0.900 (0.896, 0.904)	0.904 (0.900, 0.908)	0.904 (0.900, 0.908)	0.904 (0.900, 0.908)	**0.908*** **(0.904, 0.912)**)
	F1-Macro	0.597 (0.590, 0.603)	0.617* (0.610, 0.623)	0.620* (0.614, 0.626)	0.610 (0.603, 0.616)	**0.625*** **(0.619, 0.631)**

Sufficiency of >0.2 achieved the best performance improvements across all tested thresholds. Sufficiency of >0.817 represents the 90% percentile of rationales with false predictions for falsely predicted reports. The 95% confidence interval is computed via normal approximation [115].

(regimen iii), the macro F1-score did increase by 3%. In contrast, the CLF-DPLA achieved significant performance improvements when trained on (ii) versus (i). The achievable performance improvements for the CLF-DPLA trained on sufficient rationales (regimen iii) are inconsistent—the accuracy score remains unchanged, the macro-averaged score for the sufficiency threshold of 0.2 slightly increases, and the macro-averaged score for the threshold of 0.817 decreases.

4.4.2 Class-Wise Model Performance in Low-Resource Scenario for 10 Common Primary Cancer Sites

To study how integrating rationales impacts class-wise model performance in low-resource scenarios, we simulated these scenarios by adding (to the training reports) a small, randomly selected subset of $m \in [100; 500; 1,000]$ rationales for 10 common primary cancer sites. This enabled us to test the potential benefit of adding rationales when data is scarce. From the results in Table 4.3, we can surmise that (i) the addition of rationales may lead to improved class-wise performance, (ii) sufficient rationales do not show consistent positive impacts on the class-wise performance, (iii) the number of already available reports for a given class in the training dataset (infrequent vs. frequent classes) has no consistent impact on the achievable class-wise performance, and (iv) the class-wise performance improvements do not follow a consistent pattern (e.g., multiple classes experience similar good performance with only 100 and 1,000 added rationales while having a reduced performance with 500 additional rationales), see kidney (C64). To provide a full picture of the performance of our tested models in low-resource scenarios with limited availability of rationales, we also report their overall performance in Table B.5 in subsection B.4. Generally, adding a few rationales reduces overall performance for both models. This trend persists in scenarios in which the training data is augmented with sufficient rationales.

Table 4.3: Performance improvements of the CLF-BS and CLF-DPLA models when classifying primary cancer sites from SEER pathology reports. Class-wise F1-scores for 10 common cancer sites are shown as a function of $m \in [100, 500, 1,000]$ additional rationales supplemented to the original reports dataset (n=87,252). HRS is hematopoietic and reticuloendothelial systems.

ICD-O-3 Primary Cancer Site		Stomach (C16)			Colon (C18)			Pancreas (C25)			HRS (C42)			Skin (C44)		
Number of Training Reports		1,612			3,679			1,849			7,636			5,646		
Number of Training Rationales		1,728	1,617	1,275	4,145	4,044	3,164	2,085	1,905	1,459	8,122	8,089	7,418	6,349	6,204	5,670
Ratio of Training Rationales		1.0	0.94	0.74	1.0	0.98	0.76	1.0	0.91	0.70	1.0	1.0	0.91	1.0	0.98	0.89
Model	Number of Added Rationales	Sufficiency			Sufficiency			Sufficiency			Sufficiency			Sufficiency		
		> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817
CLF-BS	0	0.789	–	–	0.847	–	–	0.802	–	–	0.904	–	–	0.961	–	–
	100	0.787	0.782	0.770	0.845	0.853	0.854	0.800	0.808	0.803	0.909	0.900	0.908	0.957	0.961	0.963
	500	0.781	0.772	0.783	0.847	0.847	0.851	0.807	0.808	0.811	0.904	0.892	0.900	0.964	0.955	0.962
	1,000*	0.766	0.772	0.773	0.854	0.853	0.854	0.812	0.816	0.801	0.898	0.901	0.901	0.959	0.962	0.962
CLF-DPLA	0	0.813	–	–	0.898	–	–	0.826	–	–	0.908	–	–	0.971	–	–
	100	0.788	0.798	0.804	0.893	0.891	0.894	0.827	0.816	0.822	0.907	0.908	0.911	0.971	0.971	0.974
	500	0.804	0.796	0.797	0.896	0.877	0.888	0.824	0.836	0.834	0.906	0.909	0.910	0.973	0.970	0.974
	1,000*	0.795	0.806	0.781	0.889	0.887	0.894	0.812	0.827	0.819	0.902	0.901	0.903	0.969	0.969	0.970
ICD-O-3 Primary Cancer Site		Breast (C50)			Corpus uteri (C54)			Prostate gland (C61)			Kidney (C64)			Thyroid gland (C73)		
Number of Training Reports		25,160			3,521			1,847			1,068			1,500		
Number of Training Rationales		28,408	28,196	27.017	3,790	3,615	2,925	1,964	1,865	1,682	1,134	1,060	973	1,573	1,509	1,455
Ratio of Training Rationales		1.0	0.99	0.95	1.0	0.95	0.77	1.0	0.95	0.86	1.0	0.93	0.86	1.0	0.96	0.92
Model	Number of Added Rationales	Sufficiency			Sufficiency			Sufficiency			Sufficiency			Sufficiency		
		> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817	> 0	> 0.2	> 0.817
CLF-BS	0	0.982	–	–	0.907	–	–	0.816	–	–	0.521	–	–	0	–	–
	100	0.980	0.980	0.980	0.907	0.900	0.907	0.829	0.818	0.806	0.522	0.518	0.479	0.226	0.048	0.175
	500	0.981	0.983	0.981	0.909	0.902	0.910	0.845	0.830	0.833	0.554	0.553	0.553	0.222	0.165	0.044
	1,000*	0.982	0.981	0.979	0.904	0.901	0.899	0.836	0.836	0.831	0.432	0.541	0.539	0.089	0.217	0.044
CLF-DPLA	0	0.984	–	–	0.934	–	–	0.868	–	–	0.511	–	–	0.307	–	–
	100	0.986	0.986	0.985	0.927	0.927	0.928	0.864	0.855	0.854	0.477	0.570	0.564	0.180	0.039	0.122
	500	0.986	0.987	0.987	0.927	0.918	0.922	0.841	0.861	0.876	0.503	0.504	0.537	0.313	0.276	0.377
	1,000*	0.987	0.987	0.984	0.922	0.927	0.928	0.871	0.859	0.855	0.521	0.575	0.542	0.152	0.140	0.089

4.4.3 Human-based Clinical Rationales and Model Explainability

Next, we analyze the potential impact of rationales on a model's explainability from the perspective of attention. For this analysis, we focused on the better performing CLF-DPLA. Because the DPLA deploys two target attention mechanisms, we can use the generated attention scores* as a gauge for the relative importance of a token at the token and phrase levels for the primary cancer site classification task. For all 2,446 rationale-annotated test reports, we calculated the overlap ratio between the rationale tokens and the tokens with the highest attention scores by using the 90, 95, and 98 attention score percentiles as thresholds. Notably, because transformers utilize a byte-pair encoded vocabulary with subwords and we want to represent the ratios at the word level, we recreated the word-level attention scores as the average of all subword-level scores. We hypothesize that the type of training input impacts the token-level rationale coverage ratios, so we computed the ratios for models trained on only rationales, only complements, only reports, reports supplemented with all rationales, and reports supplemented with additional labeled reports.

Figure 4.1 presents the token-level rationale coverage ratio averaged across all the test documents for the CLF-DPLA and indicates that the ratios are impacted by the selected attention score percentile and the type of training data. For both attention mechanisms, models trained on rationales achieved the lowest overlap, whereas models trained on reports and rationales achieved the highest overlap with the available rationale annotations. Interestingly, supplementing the training data with additional reports enables models to achieve similar high overlap with rationales for the token-level attention mechanism but only medium overlap for the phrase-level target attention mechanism. In contrast, models trained on only reports or complements had a medium overlap with the rationale annotations associated with the test documents. Appendix B.4 provides a more detailed breakdown of the average

*We acknowledge the controversy and discussion about the suitability of attention scores for providing sufficient model explainability [19].

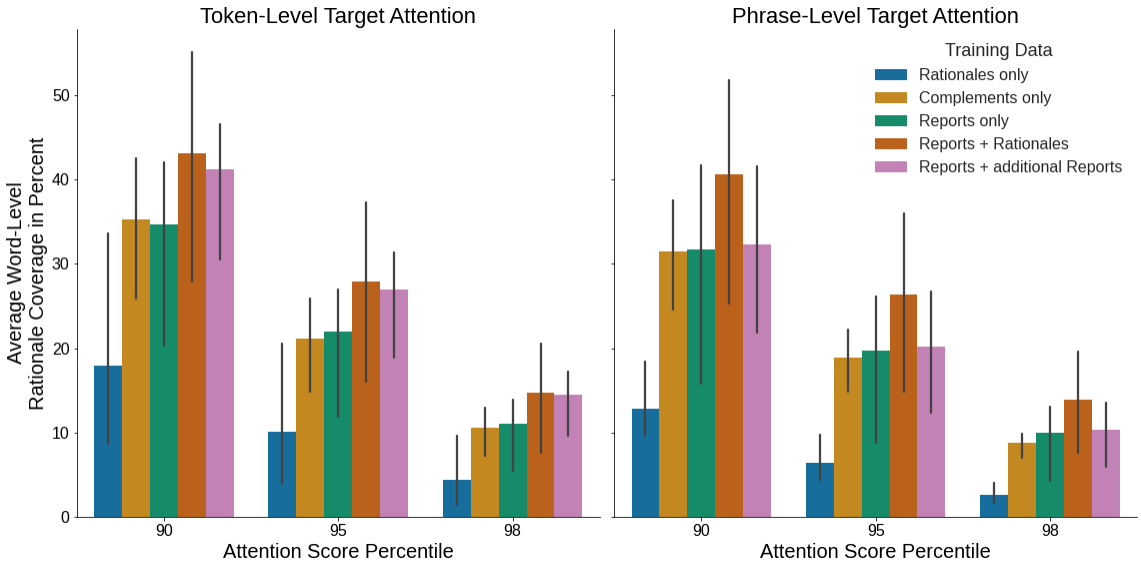


Figure 4.1: Average token-level rationale coverage across different attention score percentiles for DPLA's word- and phrase-level target attention mechanisms. Rationale coverage is calculated as the proportion of tokens deemed highly important (based on attention scores) that overlap with human-annotated rationales. Results are aggregated from 2,446 test electronic pathology reports. The error bars represent the standard deviation between the average token-level rationale coverage ratios across different random initialization seeds.

token-level rationale coverage ratios by prediction (i.e., correct versus false), rationale length, and training regimen–specific overlap of the highlighted tokens. In short, the rationale overlap (i.e., average token-level rationale coverage) is higher for correct predictions across all training regimens, tend to follow a unimodal distribution across sequence length with a peak coverage for sequences ranging from 6 to 15 words, and the different training regimens highlight up to 60% of the same tokens associated with the rationales.

4.5 Discussion

4.5.1 Impact of Human Expert–Derived Rationales on Model Performance

We explored the impact of rationales on clinical text classification models used for automated medical encoding of electronic pathology reports with primary cancer site diagnoses. Our experiments revealed that rationales can positively influence model performance. However, we observed decreased overall and inconsistent class-wise performance in low-resource scenarios, a finding that suggests potential issues with rationale quality. Interestingly, these inconsistencies in model performance remain for models trained on sufficient rationales. Lastly, the increase in classification performance from augmenting the training data with rationales was smaller than the increase from our control model that simply trained on more pathology reports.

Although our findings align with previous research on enhancing ML model performance using rationales [9, 162, 116] in high-resource scenarios, the observed improvements in accuracy are somewhat underwhelming in terms of the annotation effort required to achieve performance increases given the extensive addition of rationales—effectively doubling the size of the training dataset. This is particularly notable when compared to the minimal improvements over the baseline performance (i.e., models trained on the reports without rationales) and the lower performance

compared with the control models trained on additional labeled reports. Therefore, we think that the rather modest improvements achieved with supplementing rationales over the models trained on only reports and their low performance compared with models trained on a larger volume of pathology reports raises concerns about the return on investment for annotating rationales. Essentially, when improving the model accuracy is the goal, we recommend allocating resources to acquiring and annotating new, unlabeled samples instead of annotating rationales in already collected and labeled documents. Notably, we agree with [158] observation that the simultaneous annotation of the ground-truth label (i.e., diagnosis) and the rationale maximizes the return on investment during data annotation.

Given the considerable effort required to annotate rationales and the low availability of datasets with a large number of annotated rationales [146], it is important to evaluate model performance in low-resource scenarios. Contrary to our expectations, the addition of a few rationales—randomly sampled from 10 common primary cancer sites—to the training data caused a drop in overall model performance and an interesting display of inconsistent trends in class-wise performance. The drop and inconsistent trends in performance may be caused by unwanted class-wise interactions between the introduced rationales, general quality-related issues with the set of available rationales, or rationales introducing signals that possibly contradict the information a model finds useful in the original training reports for a given class; this information may stem from spurious correlations or artifacts introduced through the data [53]. Furthermore, the mismatch in signal between rationales and reports has been described by [116] as data distribution shift possibly causing a decrease in model performance. This data distribution shift could also explain the inconsistent model performances in high- versus low-resource scenarios, in which models may experience the introduction of rationales as noise in low-resource scenarios because the class-wise input features provided by the rationales differ from the features provided by the reports of the same class. Although the analysis of possible interactions between classes is worthwhile, it is beyond the scope of this study.

Another approach to explain the mismatch in signal between rationales and reports and the potential quality-related issues can be borrowed from the explainable AI literature. In this space, rationales are seen as explanations [136], which can be faithful (i.e., able to explain a model's inner decision making) and/or plausible (i.e., understandable by humans). Although we think our rationales provide highly plausible explanations because they were derived by subject matter experts—acknowledging typical variations in annotation quality—our rationales may not necessarily provide faithful information.

In our first analysis, we evaluated the faithfulness of our rationales by comparing the model performance when trained on only rationales, on full reports, or on complements. According to the literature, faithful rationales should ideally enable models to perform comparably to those trained on full reports [91]. However, our experiments found that models trained on only rationales achieved the poorest performance of the group, whereas report-trained models achieved the highest performance (Table 4.4). These findings suggest that (i) ML models rely on a broad range of information presented in clinical documents to optimally learn the decision boundary space; (ii) plausible explanations (i.e., text that a human would find crucial for a clinical diagnosis) may not necessarily be as important to a ML model as they are to humans; (iii) ML models may find seemingly irrelevant information, possible spurious correlations, or data artifacts to be critical in developing their decision boundary space; (iv) cancer pathology reports are complex and contain multiple crucial concepts that a model potentially relies on during its decision-making; and (v) to a model, rationales represent only a single source of crucial information provided by the cancer pathology reports. Additionally, rationales appear to provide essential information in reports associated with minority classes indicated by the reduced F1-macro scores of models trained on complements compared with models trained on reports. Our findings align with the assessment of Carton et al., who argued that models are susceptible to biases through artifacts in the data [26], thereby causing inconsistent or false evaluation of rationales, and who empirically observed that

Table 4.4: Comparison of accuracy and F1-macro scores for different training input types for the base CLF-BS and the CLF-DPLA when performing the classification of the primary cancer site in SEER cancer pathology reports trained on only the reports ($n = 87,252$), rationales ($n = 96,679$), or the rationale complements ($n = 96,679$). The models were tested on the training split containing the reports, validation ($n = 19,385$), and test splits ($n = 22,012$).

Model	Split	Metric	Training Data Input		
			Rationales	Complements	Reports
CLF-BS	Train	Accuracy	0.840	0.916	0.932
		F1-Macro	0.468	0.644	0.707
	Val	Accuracy	0.802	0.878	0.880
		F1-Macro	0.437	0.537	0.552
	Test	Accuracy	0.801	0.881	0.885
		F1-Macro	0.437	0.553	0.567
CLF-DPLA	Train	Accuracy	0.691	0.916	0.937
		F1-Macro	0.328	0.670	0.755
	Val	Accuracy	0.793	0.900	0.901
		F1-Macro	0.433	0.583	0.600
	Test	Accuracy	0.780	0.900	0.900
		F1-Macro	0.426	0.579	0.597

human rationales may not provide sufficient information for the model to exploit for prediction [25]. Lastly, we are not surprised that models trained on only rationales performed far worse than their report- and complement-trained counterparts given that our rationales (on average) cover only roughly 4% of the full text and many rationales consist of only a few words.

Our second analysis followed recent work that evaluated rationale quality by decomposing faithfulness into two key properties: sufficiency and comprehensiveness [37, 26]. Sufficiency attempts to measure how well a rationale can substitute the full information to reproduce the prediction, whereas comprehensiveness assesses how well a rationale encapsulates all information relevant for the prediction. Therefore, a high-quality faithful rationale should have both high sufficiency and high comprehensiveness [26]. Furthermore, in their comprehensive survey on datasets with annotated rationales, Wiegreffe et al. recommended avoiding the use of low-sufficiency rationales and to assess whether the collected rationales truly explain the ground-truth label (i.e., primary cancer site) [146]. We computed the sufficiency

and comprehensiveness scores by using the prediction probabilities generated from the experiments that compared model training on rationales, full reports, and complements. The scores are shown in Figure 4.2.

Generally, our rationales achieve sufficiency scores similar to those in prior published work on non-clinical, human-based rationales. However, the comprehensiveness scores for our rationales are lower than the published ones [26]. The low comprehensiveness scores further support the claim made above: the rationales, although containing crucial information for a given report's diagnosis, only cover a small fraction of the relevant information in the full report.

Interestingly, both metrics are impacted by the support for each class in the dataset, with sufficiency increasing and comprehensiveness decreasing with an increased number of samples for a given class. It is unclear from our set of experiments whether this trend for both metrics is caused by the actual quality associated with

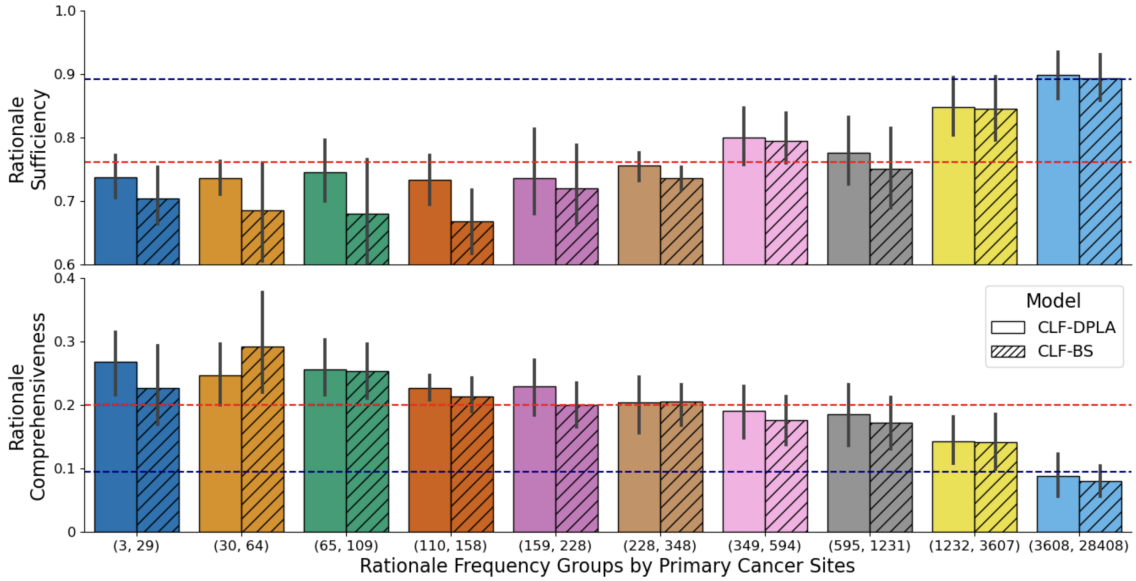


Figure 4.2: Evaluation of rationale faithfulness decomposed into sufficiency and comprehensiveness for the CLF-BS and CLF-DPLA models for all evaluated primary cancer sites grouped by their frequency occurrence. The blue-dashed line depicts the mean score averaged across all samples, and the red-dashed line is the mean score averaged across all classes.

the rationales across classes or by having more class-wise support, thereby allowing the ML-based model to become more confident in its predictions. Given that both metrics are based on prediction probabilities, we conclude that the latter case is more likely, especially considering that each rationale was created for a specific cancer pathology report and primary cancer site diagnosis. Although minority classes occur infrequently, we do not anticipate that label frequency will significantly impact the rationale annotation process or substantially degrade rationale quality.

Finally as a third analysis, we wondered if the sufficiency metric could be used as a preselection threshold to determine higher-quality rationales to improve model performance (see Tables 4.3 and B.5). Generally, preselecting rationales based on sufficiency scores does not consistently improve performance when tested across the full set of rationales in low-resource scenarios. Although we see improved overall performance when augmenting the training dataset with *sufficient* rationales, the performance in low-resource scenarios is inconsistent, with performance dropping for colon (C18) and for thyroid gland (C73), indicating that sufficiency alone is not an ideal metric for rationale preselection.

4.5.2 Impact of Human Expert–Derived Rationales on Model Explainability

Next, we discuss the potential impact of our rationales on model explainability. Explainability is understood as a model's ability to provide plausible explanations for its predictions (i.e., the greater the overlap between features highlighted by the model as important and human annotated rationales, the more plausible the model [146]). Model plausibility is especially important in medicine. Similar to [116], we hypothesized that supplementing the training regimen with plausible explanations in the form of our rationales can teach a model to focus its attention on more rationale-like features. We computed the average token-level rationale coverage for 2,446 test documents using attention scores generated by the CLF-DPLA when trained on

reports supplemented with rationales. We then contrasted those results with reports only, rationales only, complements only, and reports supplemented with additional reports (Figure 4.1).

Our analysis shows that the type of information used for model training influences rationale coverage, and this indicates that model plausibility can be impacted by the selection of the training data. The results further indicate that rationales alone are not a sufficient source of information to *teach* models plausibility. In particular, models do not necessarily need the rationale information, and models trained on complements or full reports achieve similar average token-level rationale coverage. The analysis shows that the different types of training input result in an overlap of highlighted rationales of up to 60% (see Appendix B.4). It is only when we combine rationales with full reports that rationales are helpful for teaching models to look for rationale-like features. Furthermore, our results also show that statistically similar levels of plausibility can be achieved through supplementing the training data with additional reports.

Although we think that rationales that encapsulate plausible explanations for a primary cancer site diagnosis can be used to teach models to identify rationale-like features as important, we strongly encourage annotators to ensure that the extracted rationales contain sufficient information from the report without introducing noise through the addition of words not relevant to the actual diagnosis. Appendix B.4 shows that the average token-level rationale coverage is impacted by rationale length. Although our experiments show that rationales may not consistently improve model performance, rationales can positively impact model explainability. Our findings suggest that clinical rationales serve as an effective source of information for priming AI models to generate plausible explanations while maintaining robust performance. This balance between explainability and accuracy is crucial for the successful implementation of AI-driven clinical support systems. Therefore, model optimizations must focus on both aspects to enhance the trustworthiness of such systems and increase their acceptance among medical professionals [162, 116]. Interestingly, our

experiments showed that this trade-off exists, as highlighted by the improved rationale coverage and relatively consistent performance for models trained on rationales.

4.6 Conclusion, Limitations, and Future Work

We investigated the impact of integrating human-based clinical rationales into training data to improve the performance of transformer-based clinical text classification models used to classify SEER electronic pathology reports with the primary cancer site diagnosis. Our results show that adding rationales can improve model performance in high-resource scenarios but achieves inconsistent performance in low-resource scenarios. Additionally, rationale-trained models are outperformed by models trained on a dataset augmented with more reports. Our results highlight the importance of measuring and ensuring rationale faithfulness with appropriate automatic metrics because any quality-related issues can degrade model performance.

The main conclusions of this work are as follows: (i) If improving model performance is the goal, then the highest return on investment for data annotations is achieved by annotating additional documents instead of extracting rationales from already available reports. (ii) For scenarios in which additional reports with full information are unavailable, the annotation of all available reports with rationales can effectively improve a model's overall performance. However, the selective annotation for frequent classes may not consistently improve class-wise performance; the selective annotation of minority classes should be considered. (iii) Utilizing automatic metrics (e.g., sufficiency) to assess a rationale's quality leads to inconsistent improvements in model performance. (iv) Finally, including rationales in the training data can improve model explainability by priming a model to pay more attention to rationale-like subsequences in a document during inference.

Although this study offers valuable insights for using rationales to improve clinical text classification models, one must consider its limitations and potential avenues for future research. A major limitation of this study is the introduction of a

sampling bias that occurs when collecting data from the reports annotated with rationales. This sampling bias is expressed by creating rationales for reports that failed the initial quality control and a registry-specific data origin check. Another limitation is that our study focuses primarily on the evaluation of extractive rationales to improve discriminative models and neglects abstractive rationales created by generative models. Future research could explore the strengths of LLMs to either create *synthetic* extractive rationales or investigate methods that force LLMs to self-rationalize when classifying pathology reports with the primary cancer sites. Another direction could be the investigation of automatic metrics that appropriately measure the quality of rationales.

Chapter 5

Conclusion

5.1 Summary of Findings

This dissertation is centered around the following research questions:

1. How does the reference information in popular attention mechanisms impact model performance? Does the initialization strategy of the reference information matter? Can you encode the information on the code hierarchical of medical encoding systems via the reference information?

2. Can traditional word-level attention mechanisms be extended to accurately extract complex clinical language from clinical text by identifying longer, cohesive groups of words or phrases to improve the performance and explainability of clinical text classification models?

3. Can human-based clinical rationales improve the performance and explainability of attention-based clinical text classification models?

Chapter 2 addressed the first research questions by analyzing the role of the reference information contained by the query matrix of modern attention mechanisms, particularly the label-wise attention mechanism proposed by Mullenbach et al. (2018) [99]. The label-wise attention mechanism is the most frequently used form of attention in state-of-the-art multi-label clinical text classification models. In particular, this study evaluated the impact of the reference information used by the label-attention mechanism across four text-encoder architectures—convolutional neural networks, bi-directional gated recurrent neural network, bi-directional long short-term memory neural networks, and the transformer-based clinical longformer—on the automated medical encoding of MIMIC-III hospital discharge summary notes with ICD-9 billable codes. The results are contrasted against the simpler target-attention mechanism and common methods to generate document embeddings.

The experiments revealed two key findings. First, the type of initialization of the reference information with either random or pretrained weights significantly

impacts model performance. Second, attention mechanisms can learn to encode the hierarchical structure inherent in medical coding systems through a strategic design of the information flow. This capability lead to improved performance in clinical text classification models for the automated medical encoding of hospital discharge summary notes. Notably, incorporating hierarchical information through attention mechanisms eliminates the need for training a separate, auxiliary neural network, resulting in a streamlined model architecture.

Chapter 3 addressed the second research question by proposing the novel deformable phrase-level attention mechanism (DPLA). The DPLA was designed to dynamically capture the most important word-level lexical and phrase-level contextual information in a clinical document. The DPLA is an independent module usable to extend any deep learning text-encoder architecture. The DPLA was tested on a multi-class and multi-label clinical text classification tasks. The first set of experiments focused on the automated information extraction of critical cancer data elements—site, subsite, laterality, histology, and behavior—from 630k SEER electronic pathology reports. The second set of experiments focused on the automated medical encoding of MIMIC-III hospital discharge summary notes with multiple billable ICD-9 codes.

The results showed that the DPLA is effective in capturing lexical word-level and contextual phrase-level information from clinical documents improving the performance of clinical text classification models on both classification tasks. In particular, extending the transformer-based Clinical Longformer with the DPLA significantly outperformed all baseline models—including the previous state-of-the-art hierarchical self-attention network (HiSAN) model—and achieved the highest performance across all oncological tasks. In contrast, the experiments on the MIMIC-III data showed that all three evaluated text-encoder architectures (CNN, Bi-GRU, and transformer-based CLF), when extended with the DPLA, showed equal or better performance compared with the popular word-level, label-wise attention mechanism [99].

Chapter 4 addressed the final research question by exploring the suitability of human-based clinical rationales (HBCR) as additional supervision signals to improve the performance and interpretability of clinical text classification models. Multiple sets of experiments were performed to gauge the achievable model performance when trained on data supplemented with HBCR in high- and low-resource scenarios—high-resource scenarios assumes to have access to HBCR across all labels, whereas, low-resource assumes to only have access to HBCR of a few classes. The achievable performance improvements through HBCR were contrasted against models trained on a dataset supplemented with a subset of *sufficient* HBCR filtered using the *sufficiency* metric [37, 26] and models trained on a dataset supplemented with additional electronic pathology reports. The experiments focused on the information extraction of the primary cancer site from electronic pathology reports.

Our experiments revealed mixed behavior regarding the impact of human-based clinical rationales (HBCR) on model performance. While performance improves with abundant HBCR, results are inconsistent in low-resource scenarios. Moreover, models trained on datasets supplemented with more additional electronic pathology reports outperformed those trained supplemented with rationales. This finding suggests that if improving model performance is desired, than the annotation efforts should prioritize labeling additional documents over extracting rationales from existing reports. While improving model performance is general the primary goal in improving text classification models, in medical applications like ours improving model interpretability is equally important. Analyses on model interpretability showed that HBCR have the ability to prime a model to focus on rationale-like sequences within a report during inference. These insights highlight the complex relationship between rationales and model performance, suggesting careful consideration of annotation strategies and quality assessment in clinical NLP tasks.

5.2 Impact and Implications

Attention mechanisms are a crucial component in contemporary state-of-the-art clinical text classification models. These models typically utilize a variety of attention mechanisms (e.g., target-, label-, or self-attention) that aim to capture the essential aspects of clinical text relevant to the classification task at the word level. While word-level attention mechanisms help models identify global relationships between words based on the intrinsic characteristics of the data, they often neglect the local semantic information crucial for understanding the complex and ambiguous language inherent to clinical text [122]. To overcome this barrier, and improve the understanding of clinical text, this dissertation in part designed novel attention mechanisms to either utilize auxiliary information related to the task (i.e., code hierarchy or textual code description) (see chapter 2) or employ an internal mechanism designed to accurately extract contextual information at the phrase level (see chapter 3).

The work in chapter 2 showed that attention mechanisms can be initialized with explicit reference information on the label space (e.g., class descriptions or information on the code hierarchy inherent to medical encoding systems) in order to prime the model to search for similar passages in the clinical documents. This is essential for DL models, as this gives ML practitioners an opening to control what the model should look for in the input data. Providing such information can help move a model's internal decision making closer to that of a human expert, which would increase the trustworthiness of such AI-driven clinical support systems. In particular, subject matter experts in the clinical domain are encouraged to develop sufficient and comprehensive reference information which describe the unique properties of a medical code (i.e., disease) in order to teach the model to discriminate between diagnoses closer to the level of human experts.

The work in chapter 3 demonstrated how harvesting local contextual information at the phrase level via the deformable phrase-level attention mechanism (DPLA) better informs a model on the important aspects of a clinical document. This

approach led to significantly better performance over the current state-of-the-art information extraction model, multi-task hierarchical self-attention network (MT-HiSAN) (see Table 3.1). Particularly noteworthy are the performance improvements on the notoriously difficult tasks of *subsite* and *histology*, where previous models struggled with predicting the correct classes due to the large and imbalanced label space. This flexible mechanism is particularly important, as this flexibility allows the model to mirror human comprehension more closely by evaluating smaller or larger contextual windows for different.

Another benefit of the proposed DPLA is that models generate a smaller proportion of documents with uncertain predictions in high-risk scenarios where only a small error rate is tolerated (i.e., 3%). As Table 3.2 shows, the CLF-DPLA rejects 4% to 9% fewer documents compared to the MT-HiSAN across all five information extraction tasks—site, subsite, laterality, histology, and behavior. In clinical practice, this result implies that pathologists need to annotate up to 9% fewer documents than with prior models, which is significant considering that up to two million new cancer pathology reports are created annually. Consequently, pathologists' time is freed up, allowing them to perform more critical tasks such as personalized cancer care.

A limitation of the DPLA is that it could not reliable provide plausible explanations for its diagnosis predictions. Providing reliable plausible explanations, that are also of causal nature, to human experts is critical for building trust and enthusiasm of medical professionals for adopting AI-driven clinical support systems in their daily practice. In order to help improve a models performance and explainability, previous work in the non-medical domain has explored human-based rationales.

The research presented in Chapter 4 investigated the use of human-based clinical rationales (HBCR) as supplementary training data to enhance both performance and interpretability of clinical text classification models. Contrary to previous studies which demonstrated the effectiveness of rationales in improving machine learning-based text classification for simpler tasks, the results of our experiments showed that HBCR inconsistently improve model performance. One possible reason for this

discrepancy is that previous work applied rationales on binary text classification tasks with simpler text documents (e.g., movie reviews [158]). Whereas clinical text classification are much harder due to the presence of large label spaces the difficult nature of clinical text documents. In particular, clinical text documents are particularly challenging due to their use of complex clinical language and the possible existence of multiple contextual topics making learning from HBCR much more difficult. It is therefore paramount, that annotators provide sufficient and comprehensive clinical rationales that contain document-specific information relevant to the ground-truth label, but are also distinctly different at the class-level. This balance, however, is expected to be difficult due to the complexity of clinical documents and the similarity between classes (i.e., diagnoses).

Another significant finding was that models trained on a dataset supplemented with HBCR were outperformed by models trained on simply more labeled reports which raised concerns about the return on investment for data annotation. Essentially, for scenarios like ours where improving the model performance without introducing data bias is the goal, we recommend to allocate resources towards acquiring and annotating new, unlabeled samples instead of delayed annotation of rationales in already collected and labeled documents. However, if the collection of clinical rationales is desired, it is generally recommended to simultaneously annotate rationales and the ground-truth labels for new, unlabeled reports to increase annotation efficiency. Zaidan et al. [158] pointed out that annotators naturally need to identify the rationales when assigning ground-truth labels to documents, making explicit rationale annotation a complementary and relatively straightforward task resulting in increased annotation efficiency.

Lastly, the results showed that HBCR have the capability to prime clinical text classification models to highlight rationale-like features as important without severely degrading model performance. This finding is promising, since in medicine, model optimization focuses on the improvement of both the performance and explainability, as opposed to non-medical domains which focus solely on improving

model performance. Developing novel methods that can improve model performance and explainability which will lead to increased trustworthiness of medical professionals in AI-driven clinical support systems. While this work showed that models can be primed in finding rationale-like features, this work did not provide a method that can generate causal explanations leaving an important avenue for future work.

5.3 Future Work

This dissertation focused on developing novel methods to primarily improve the performance of clinical text classification models performing the automated medical encoding of clinical text documents. While some of the methods showed favorable properties to provide an avenue for interpretation of a model's predictions, the proposed methods did not explicitly focus on improving model explainability. Model explainability is defined as the provision of an output that either faithfully explains the inner workings of a model's decision making (i.e,. provides causality) or plausibly explains a model's prediction understandable to a human target audience [91]. In our application, this target audience are medical professionals that are expected to utilize and rely on AI-driven clinical support systems in their daily practice. An essential, non-negotiable property of these systems is their trustworthiness; the ability of medical professionals to assess the validity, reliability, and limitations of AI-driven clinical support systems. Therefore, it is crucial future work that the improvement of AI-driven clinical support systems must follow a multi-object optimization strategy which simultaneously targets improving model performance and trustworthiness. Hence, the following questions should be seen as a starting point for applied future work:

- How do you efficiently optimize for trustworthiness?
- What are the metrics used to measure model trustworthiness?
- Which specific aspects links model trustworthiness to explainability?

- What are the metrics used to measure model explainability? What is the relationship between faithful and plausible explainability in terms metric utilization? Is a model's ability to retrieve rationale-like features as measured by the overlap of human-based clinical rationale ground-truth labels sufficient or are non-annotated rationale-like features also important?

- How do you link model reasoning to model explainability? Which role does causality play in model explainability?

- Do the models provide *learned* explanations or provide intuitive reasoning not explicitly learned during model training?

- Can you trust connections that are seemingly not plausible (e.g., spurious associations or data artifacts)? Which metrics or proof are required to satisfy medical professionals?

- Does the required level of model trustworthiness depend on the specific application in medicine? For example, automated medical encoding of clinical text documents for the population of databases vs. direct clinical practice such AI-driven automated cancer detection in images?

Bibliography

[1] Cdc population level surveillance. https://www.cdc.gov/surveillance/data-modernization/index.html#print. Accessed: 2024-03-06. 67

[2] National Cancer Institute seer. https://seer.cancer.gov/data-software/. Accessed: 2022-07-05. 6, 59, 86, 139

[3] Jul 2023. 59

[4] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022. 13

[5] Z. Ahmad, S. Rahim, M. Zubair, and J. Abdul-Ghafar. Artificial intelligence (ai) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review. *Diagnostic pathology*, 16:1–16, 2021. 48

[6] M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphrey, X.-C. Wu, L. Coyle, and G. Tourassi. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. *Journal of the American Medical Informatics Association*, 27(1):89–98, 2020. 14

[7] M. Alawad, H.-J. Yoon, and G. D. Tourassi. Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 218–221. IEEE, 2018. 67, 143

[8] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of

deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021. 8

[9] I. Arous, L. Dolamic, J. Yang, A. Bhardwaj, G. Cuccu, and P. Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876, 2021. 15, 85, 97

[10] M. Ayaz, M. F. Pasha, M. Y. Alzahrani, R. Budiarto, and D. Stiawan. The fast health interoperability resources (fhir) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR medical informatics*, 9(7):e21929, 2021. 57

[11] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 12

[12] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016. 20

[13] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, et al. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88, 2019. 3

[14] S. Banerjee, C. Akkaya, F. Perez-Sorrosal, and K. Tsioutsiouliklis. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, 2019. 23

[15] Y. Bao, S. Chang, M. Yu, and R. Barzilay. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*, 2018. 85

[16] J. Bastings and K. Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*, 2020. 81

[17] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 11, 26

[18] Y. Bengio, R. Ducharme, and P. Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000. 10

[19] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, 2022. 95

[20] A. Blanco, S. Remmer, A. Perez, H. Dalianis, and A. Casillas. Implementation of specialised attention mechanisms: Icd-10 classification of gastrointestinal discharge summaries in english, spanish and swedish. *Journal of Biomedical Informatics*, 130:104050, 2022. 14

[21] W. Blumenthal, T. O. Alimi, S. F. Jones, D. E. Jones, J. D. Rogers, V. B. Benard, and L. C. Richardson. Using informatics to improve cancer surveillance. *Journal of the American Medical Informatics Association*, 27(9):1488–1495, 2020. 2, 3, 57

[22] G. Brauwers and F. Frasincar. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 12, 27

[23] J. M. Buckley, S. B. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A. K. Belli, E. M. Kim, J. E. Garber, B. L. Smith, M. A. Gadd, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *Journal of pathology informatics*, 3(1):23, 2012. 3

[24] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, 2020. 14, 21, 23, 36

[25] S. Carton, S. Kanoria, and C. Tan. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088, 2022. 100

[26] S. Carton, A. Rathore, and C. Tan. Evaluating and characterizing human rationales. *arXiv preprint arXiv:2010.04736*, 2020. 4, 15, 85, 86, 90, 99, 100, 101, 109, 155

[27] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. *arXiv preprint arXiv:2010.01653*, 2020. 21

[28] A. Chan, Z. Zeng, W. Lake, B. Joshi, H. Chen, and X. Ren. Knife: Distilling meta-reasoning knowledge with free-text rationales. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 15, 85

[29] H. Chen, Q. Ma, Z. Lin, and J. Yan. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4370–4379, 2021. 23, 51

[30] M. Chen, K. Q. Weinberger, F. Sha, et al. An alternative text representation to tf-idf and bag-of-words. *arXiv preprint arXiv:1301.6770*, 2013. 10

[31] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 11, 27, 59, 67, 143

[32] N. Chomsky. *Syntactic structures*. Mouton de Gruyter, 2002. 9

[33] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019. 4

[34] K. C. Costley and J. Nelson. Avram noam chomsky and his cognitive development theory. *Online Submission*, 2013. 9

[35] K. De Angeli, S. Gao, A. Blanchard, E. B. Durbin, X.-C. Wu, A. Stroup, J. Doherty, S. M. Schwartz, C. Wiggins, L. Coyle, et al. Using ensembles and distillation to optimize the deployment of deep learning models for the classification of electronic cancer pathology reports. *JAMIA open*, 5(3):ooac075, 2022. 68, 152

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 11, 13

[37] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019. 15, 86, 90, 100, 109

[38] H. Dong, V. Suárez-Paniagua, W. Whiteley, and H. Wu. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728, 2021. 13, 14

[39] J. Edin, A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, and L. Maaløe. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. *arXiv preprint arXiv:2304.10909*, 2023. 36, 76, 150

[40] M. Elhaddad and S. Hamam. Ai-driven clinical decision support systems: an ongoing pursuit of potential. *Cureus*, 16(4), 2024. 84

[41] P. Fang, W. He, D. Gomez, K. E. Hoffman, B. D. Smith, S. H. Giordano, R. Jagsi, and G. L. Smith. Racial disparities in guideline-concordant cancer care and mortality in the united states. *Advances in radiation oncology*, 3(3):221–229, 2018. 2, 7, 57

[42] S. Feldman. Nlp meets the jabberwocky: Natural language processing in information retrieval. *ONLINE-WESTON THEN WILTON-*, 23:62–73, 1999. 57, 62

[43] A. Ferrario and M. Nägelin. The art of natural language processing: classical, modern and contemporary approaches to text document classification. *Modern and Contemporary Approaches to Text Document Classification (March 1, 2020)*, 2020. 11

[44] M. Feucht, Z. Wu, S. Althammer, and V. Tresp. Description-based label attention classifier for explainable icd-9 classification. *arXiv preprint arXiv:2109.12026*, 2021. 13, 14, 22

[45] O. Firat, K. Cho, and Y. Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*, 2016. 22

[46] S. Frid, X. P. Duran, G. B. Cucó, M. Pedrera-Jiménez, P. Serrano-Balazote, A. M. Carrero, R. Lozano-Rubí, et al. An ontology-based approach for consolidating patient data standardized with european norm/international

organization for standardization 13606 (en/iso 13606) into joint observational medical outcomes partnership (omop) repositories: Description of a methodology. *JMIR Medical Informatics*, 11(1):e44547, 2023. 80

[47] C. Friedman, T. C. Rindflesch, and M. Corn. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765–773, 2013. 3, 57

[48] A. Galassi, M. Lippi, and P. Torroni. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308, 2020. 4, 12, 24, 31

[49] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, et al. Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607, 2021. 6, 34, 60, 68, 90, 144, 146, 159

[50] S. Gao, J. X. Qiu, M. Alawad, J. D. Hinkle, N. Schaefferkoetter, H.-J. Yoon, B. Christian, P. A. Fearn, L. Penberthy, X.-C. Wu, et al. Classifying cancer pathology reports with hierarchical self-attention networks. *Artificial intelligence in medicine*, 101:101726, 2019. 13, 14, 20, 32, 56, 57, 65, 67, 143

[51] S. Gao, A. Ramanathan, and G. Tourassi. Hierarchical convolutional attention networks for text classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 11–23, 2018. 14

[52] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, 25(3):321–330, 2018. 14

[53] M. Gardner, W. Merrill, J. Dodge, M. E. Peters, A. Ross, S. Singh, and N. A. Smith. Competency problems: On finding and removing artifacts in language data. *arXiv preprint arXiv:2104.08646*, 2021. 98

[54] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6:1225093, 2023. 85

[55] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 10

[56] T.-S. Heo, Y. Yoo, Y. Park, B. Jo, K. Lee, and K. Kim. Medical code prediction from discharge summary: Document to sequence bert using sequence attention. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1239–1244. IEEE, 2021. 14, 36

[57] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 11, 27

[58] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015. 80

[59] E. Hsu, H. Hanson, L. Coyle, J. Stevens, G. Tourassi, and L. Penberthy. Machine learning and deep learning tools for the automated capture of cancer surveillance data. *JNCI Monographs*, 2024(65):145–151, 2024. 153

[60] S. Hu, F. Teng, L. Huang, J. Yan, and H. Zhang. An explainable cnn approach for medical codes prediction from clinical text. *BMC Medical Informatics and Decision Making*, 21:1–12, 2021. 3, 14

[61] C.-W. Huang, S.-C. Tsai, and Y.-N. Chen. Plm-icd: automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*, 2022. 14, 36, 150

[62] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura. Medical text classification using convolutional neural networks. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, pages 246–250. IOS Press, 2017. 3

[63] J. Hutchins. The first public demonstration of machine translation: the georgetown-ibm system, 7th january 1954. *noviembre de*, 2005. 9

[64] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*, 2021. 86

[65] S. Jain, R. Mohammadi, and B. C. Wallace. An analysis of attention over clinical notes for predictive tasks. *arXiv preprint arXiv:1904.03244*, 2019. 22

[66] S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 81

[67] S. Ji, E. Cambria, and P. Marttinen. Dilated convolutional attention network for medical code assignment from clinical text. *arXiv preprint arXiv:2009.14578*, 2020. 14

[68] S. Ji, W. Sun, H. Dong, H. Wu, and P. Marttinen. A unified review of deep learning for automated medical coding. *arXiv preprint arXiv:2201.02797*, 2022. 21, 23

[69] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 7, 32, 59, 60, 67

[70] D. E. Jones, T. O. Alimi, P. Pordell, F. K. Tangka, W. Blumenthal, S. F. Jones, J. D. Rogers, V. B. Benard, and L. C. Richardson. Pursuing data modernization

in cancer surveillance by developing a cloud-based computing platform: Real-time cancer case collection. *JCO Clinical Cancer Informatics*, 5:24–29, 2021. 2

[71] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis. Examining attention mechanisms in deep learning models for sentiment analysis. *Applied Sciences*, 11(9):3883, 2021. 22

[72] R. Keshavamurthy, S. Dixon, K. T. Pazdernik, and L. E. Charles. Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches. *One Health*, page 100439, 2022. 57

[73] D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023. 9

[74] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 11, 26, 59, 67, 143

[75] T. Kocmi and O. Bojar. An exploration of word embedding initialization in deep-learning tasks. *arXiv preprint arXiv:1711.09160*, 2017. 10

[76] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019. 68, 90, 144, 159

[77] G. Kurata, B. Xiang, and B. Zhou. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 521–526, 2016. 23

[78] T. Kwon, K. T.-i. Ong, D. Kang, S. Moon, J. R. Lee, D. Hwang, B. Sohn, Y. Sim, D. Lee, and J. Yeo. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18417–18425, 2024. 15, 85

[79] M. LAWLEY and S. COLQUIST. Classification of pathology reports for cancer registry notifications. In *Health Informatics: Building a Healthcare Future Through Trusted Information: Selected Papers from the 20th Australian National Health Informatics Conference (HIC 2012)*, volume 178, page 150. IOS Press, 2012. 3

[80] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014. 28

[81] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 10

[82] F. Li and H. Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8180–8187, 2020. 14, 36

[83] I. Li, J. Pan, J. Goldwasser, N. Verma, W. P. Wong, M. Y. Nuzumlalı, B. Rosand, Y. Li, M. Zhang, D. Chang, et al. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511, 2022. 3, 20

[84] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022. 12, 27, 34, 59, 66, 67, 89, 144, 145, 146, 157, 158

[85] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022. 57, 62

[86] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017. 3

[87] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm. Hierarchical label-wise attention transformer model for explainable icd coding. *Journal of biomedical informatics*, 133:104161, 2022. 13, 14, 21, 23, 36, 38, 56, 67, 150

[88] Y. Liu, H. Cheng, R. Klopfer, M. R. Gormley, and T. Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, 2021. 13, 14, 23, 36, 67, 150

[89] H. Lu, L. Ehwerhemuepha, and C. Rakovski. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC medical research methodology*, 22(1):181, 2022. 26, 67

[90] Y. Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of biomedical informatics*, 72:85–95, 2017. 3

[91] Q. Lyu, M. Apidianaki, and C. Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–67, 2024. 99, 113

[92] Q. Ma, J. Yan, Z. Lin, L. Yu, and Z. Chen. Deformable self-attention for text classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1570–1581, 2021. 58, 62, 77, 149

[93] I. Makohon and Y. Li. Multi-label classification of icd-10 coding & clinical notes using mimic & codiesp. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021. 3

[94] C. S. Metzner, S. Gao, D. Herrmannova, J. Gounley, and H. Hanson. Deformable phrase level attention: A flexible approach for improving ai based medical coding. *Available at SSRN 4864687*, 2024. 90, 157, 158, 159

[95] C. S. Metzner, S. Gao, D. Herrmannova, E. Lima-Walton, and H. A. Hanson. Attention mechanisms in clinical text classification: A comparative evaluation. *IEEE Journal of Biomedical and Health Informatics*, 2024. xii, 66, 67, 75, 89, 160

[96] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 10, 26

[97] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021. 68, 90, 144, 159

[98] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018. 43

[99] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*, 2018. 7, 14, 21, 23, 27, 28, 32, 33, 36, 57, 59, 60, 65, 67, 107, 108, 150

[100] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011. 3, 9

[101] G. Napolitano, A. Marshall, P. Hamilton, and A. T. Gavin. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial intelligence in medicine*, 70:77–83, 2016. 3, 57

[102] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and S. Colquist. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association*, 17(4):440–445, 2010. 3

[103] P. X. Nguyen and S. Joty. Phrase-based attentions. *arXiv preprint arXiv:1810.03444*, 2018. 58

[104] Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021. 4, 27

[105] J. D. Osborne, M. Wyatt, A. O. Westfall, J. Willig, S. Bethard, and G. Gordon. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association*, 23(6):1077–1084, 2016. 3

[106] T. J. Osterman, M. Terry, and R. S. Miller. Improving cancer data interoperability: the promise of the minimal common oncology data elements (mcode) initiative. *JCO Clinical Cancer Informatics*, 4:993–1001, 2020. 57, 80

[107] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, 2020. 3

[108] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 10

[109] L. A. Pollack, S. F. Jones, W. Blumenthal, T. O. Alimi, D. E. Jones, J. D. Rogers, V. B. Benard, and L. C. Richardson. Population health informatics can advance interoperability: National program of cancer registries electronic pathology reporting project. *JCO Clinical Cancer Informatics*, 4:985–992, 2020. 2, 7, 57

[110] T. E. Potok, C. Schuman, S. Young, R. Patton, F. Spedalieri, J. Liu, K.-T. Yao, G. Rose, and G. Chakma. A study of complex deep learning networks on high-performance, neuromorphic, and quantum computers. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 14(2):1–21, 2018. 9

[111] D. Pruthi, R. Bansal, B. Dhingra, L. B. Soares, M. Collins, Z. C. Lipton, G. Neubig, and W. W. Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022. 85

[112] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019. 81

[113] W. Quan, Z. Chen, J. Gao, and X. T. Hu. Comparative study of cnn and lstm based attention neural networks for aspect-level opinion mining. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2141–2150. IEEE, 2018. 22

[114] S. Quazi. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8):120, 2022. 2, 56

[115] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018. 68, 70, 75, 90, 92, 144, 159, 162

[116] L. E. Resck, M. M. Raimundo, and J. Poco. Exploring the trade-off between model performance and explanation plausibility of text classifiers using human rationales. *arXiv preprint arXiv:2404.03098*, 2024. 15, 85, 86, 89, 97, 98, 102, 103, 155

[117] O. Richter and R. Wattenhofer. Normalized attention without probability cage. *arXiv preprint arXiv:2005.09561*, 2020. 43

[118] A. Rios and R. Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM conference on bioinformatics, computational biology and health informatics*, pages 258–267, 2015. 3

[119] P. Rodríguez, J. M. Gonfaus, G. Cucurull, F. XavierRoca, and J. Gonzalez. Attend and rectify: a gated attention mechanism for fine-grained recovery. In *Proceedings of the European conference on computer vision (ECCV)*, pages 349–364, 2018. 5

[120] W. Rösler, M. Altenbuchinger, B. Baeßler, T. Beissbarth, G. Beutel, R. Bock, N. von Bubnoff, J.-N. Eckardt, S. Foersch, C. M. Loeffler, et al. An overview and a roadmap for artificial intelligence in hematology and oncology. *Journal of cancer research and clinical oncology*, pages 1–10, 2023. 2, 56

[121] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024. 15, 85

[122] G. K. Savova, I. Danciu, F. Alamudun, T. Miller, C. Lin, D. S. Bitterman, G. Tourassi, and J. L. Warner. Use of natural language processing to extract

clinical cancer phenotypes from electronic medical records. *Cancer research*, 79(21):5463–5470, 2019. 3, 110

[123] S. Scott and S. Matwin. Feature engineering for text classification. In *ICML*, volume 99, pages 379–388, 1999. 3, 57

[124] K. Shah, H. Patel, D. Sanghvi, and M. Shah. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5:1–16, 2020. 3

[125] M. Sharma and M. Bilgic. Learning with rationales for document classification. *Machine Learning*, 107:797–824, 2018. 85, 86, 89, 155

[126] A. Sheikhtaheri, S. M. Tabatabaee Jabali, E. Bitaraf, A. TehraniYazdi, and A. Kabir. A near real-time electronic health record-based covid-19 surveillance system: an experience from a developing country. *Health Information Management Journal*, page 18333583221104213, 2022. 57

[127] J. Shen, W. Qiu, Y. Meng, J. Shang, X. Ren, and J. Han. Taxoclass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, 2021. 24

[128] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 58

[129] H. Shi, P. Xie, Z. Hu, M. Zhang, and E. P. Xing. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*, 2017. 14, 21, 22

[130] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023. 2, 3, 57, 66

[131] K. Sinha, Y. Dong, J. C. K. Cheung, and D. Ruths. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823, 2018. 3

[132] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. 10

[133] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 26

[134] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020. 80

[135] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651, 2010. 3, 20, 57

[136] J. Strout, Y. Zhang, and R. J. Mooney. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714*, 2019. 15, 85, 90, 99

[137] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. 80

[138] Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*DMS User Manual: Chapter 14 - Annotation Tasks*, 2020. Accessed: 2024-08-14. 89, 155

[139] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021. 3, 13, 57

[140] N. Taylor, L. Sha, D. W. Joyce, T. Lukasiewicz, A. Nevado-Holgado, and A. Kormilitzin. Rationale production to support clinical decision-making. *arXiv preprint arXiv:2111.07611*, 2021. 85

[141] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. 2, 56

[142] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 11, 12, 20, 27, 31, 62, 64, 141

[143] T. Vu, D. Q. Nguyen, and A. Nguyen. A label attention model for icd coding from clinical text. *arXiv preprint arXiv:2007.06351*, 2020. 14, 21, 23, 33, 36, 67, 150

[144] R. Weegar and H. Dalianis. Creating a rule based system for text mining of norwegian breast cancer pathology reports. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 73–78, 2015. 3

[145] WHO. ICD-9 Code Descriptions. https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/Downloads/ICD-9-CM-v32-master-descriptions.zip, 2013. [Online; last accessed 09-Nov-2022]. 28, 31

[146] S. Wiegreffe and A. Marasović. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*, 2021. 4, 15, 84, 85, 98, 100, 102

[147] A. E. Wieneke, E. J. Bowles, D. Cronkite, K. J. Wernli, H. Gao, D. Carrell, and D. S. Buist. Validation of natural language processing to extract breast cancer pathology procedures and results. *Journal of pathology informatics*, 6(1):38, 2015. 3, 57

[148] J. Wu, K. Tang, H. Zhang, C. Wang, and C. Li. Structured information extraction of pathology reports with attention-based graph convolutional network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2395–2402. IEEE, 2020. 14

[149] W. Wu, H. Wang, T. Liu, and S. Ma. Phrase-level self-attention networks for universal sentence encoding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3729–3738, 2018. 58

[150] Y. Wu, S. Zhao, and W. Li. Phrase2vec: phrase embedding based on parsing. *Information Sciences*, 517:100–127, 2020. 57

[151] L. Xu, S. Teng, R. Zhao, J. Guo, C. Xiao, D. Jiang, and B. Ren. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2459–2468, 2021. 23

[152] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016. 22

[153] W.-w. Yim, M. Yetisgen, W. P. Harris, and S. W. Kwan. Natural language processing in oncology: a review. *JAMA oncology*, 2(6):797–804, 2016. 80

[154] W. Yin, K. Kann, M. Yu, and H. Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017. 11

[155] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[156] M. Yu, S. Chang, Y. Zhang, and T. S. Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019. 90

[157] Y. Yu, J. Shen, T. Liu, Z. Qin, J. N. Yan, J. Liu, C. Zhang, and M. Bendersky. Explanation-aware soft ensemble empowers large language model in-context learning. *arXiv preprint arXiv:2311.07099*, 2023. 15, 85

[158] O. Zaidan, J. Eisner, and C. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267, 2007. 4, 15, 85, 86, 89, 98, 112, 155

[159] M. W. Zeghdaoui, O. Boussaid, F. Bentayeb, and F. Joly. Medical-based text classification using fasttext features and cnn-lstm model. In *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I 32*, pages 155–167. Springer, 2021. 3

[160] D. Zhang, C. Sen, J. Thadajarassiri, T. Hartvigsen, X. Kong, and E. Rundensteiner. Human-like explanation for text classification with limited attention supervision. In *2021 ieee international conference on big data (big data)*, pages 957–967. IEEE, 2021. 85

[161] Y. Zhang, I. Marshall, and B. C. Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795. NIH Public Access, 2016. 15, 85

[162] Z. Zhang, K. Rudra, and A. Anand. Explain and predict, and then predict again. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 418–426, 2021. 15, 85, 97, 103

[163] R. Zhong, S. Shao, and K. McKeown. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*, 2019. 85

[164] J. Zhou, C. Ma, D. Long, G. Xu, N. Ding, H. Zhang, P. Xie, and G. Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, 2020. 23

[165] X. Zhou, Y. Li, and W. Liang. Cnn-rnn based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(3):912–921, 2020. 3

Appendices

A Chapter 3 Supporting Information

A.1 Materials and Methods

Data

Table A.1 provides an overview of the number of evaluated documents in total and per split as well the number of unique classes per classification task.

Table A.1: Descriptive statistics for the in-distribution (ID) and out-of-distribution (OOD) Surveillance, Epidemiology, and End Results (SEER) electronic pathology reports and the MIMIC-III hospital discharge summary notes. N_L is the number of unique labels.

Dataset	N_{total}	N_{train}	N_{val}	$N_{test,ID}$	$N_{test,OOD}$	N_L
SEER Reports: Site	629,908	366,112	79,088	78,211	106,497	70
SEER Reports: Subsite	629,908	366,112	79,088	78,211	106,497	327
SEER Reports: Laterality	629,908	366,112	79,088	78,211	106,497	7
SEER Reports: Histology	629,908	366,112	79,088	78,211	106,497	644
SEER Reports: Behavior	629,908	366,112	79,088	78,211	106,497	4
MIMIC-III: Full	52,722	47,719	1,631	3,372	–	8,907
MIMIC-III: 50	11,368	8,066	1,573	1,729	–	50

Dataset: SEER Electronic Pathology Reports The National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) program collects millions of electronic cancer pathology reports to monitor the occurrence of cancer in the American population. This surveillance also informs policy decisions, including where to provide research resources in support of the eradication of the disease [2]. We obtained 2,618,994 pathology reports associated with 1,335,794 distinct Cancer-Tumor-Case (CTC) IDs from the SEER cancer registries in Kentucky, Louisiana, New Jersey, Washington, Utah, and New Mexico. We performed in-distribution (ID) experiments with data from the first five registries and out-of-distribution (OOD) experiments on New Mexico's cancer pathology reports (see Table A.2). Every pathology report has a unique CTC, and each CTC may be associated with one or more pathology reports. For each CTC, certified cancer registrars (CTRs) used all

Table A.2: Patient demographic characteristics.

	ID						OOD
Characteristic	KY	LA	NJ	WA	UT	Total	NM
Unique CTCs	56,828	51,466	56,561	84,734	21,876	271,465	51,900
Unique Records	121,688	101,486	111,773	144,809	43,655	523,411	106,497
Demographic Factors	No.						No.
Age (Years)							
0–19	226	163	196	251	197	1,033	211
20–39	1,341	1,235	1,389	2,066	1,210	7,241	1,651
40–59	8,875	7,574	9,469	12,316	4,410	42,644	8,018
60–79	29,395	27,966	30,038	41,455	10,705	139,559	28,245
80+	16,991	14,528	15,469	28,646	5,354	80,988	13,835
Sex							
Male	25,819	25,271	25,793	38,562	10,359	125,804	22,379
Female	31,009	26,195	30,768	46,172	11,517	145,661	29,581
Race							
White	52,207	36,334	45,554	73,532	20,447	228,074	45,606
Black or African American	3,846	14,318	6,674	2,458	139	27,435	776
Asian and NHPI	298	423	2,599	4,735	464	8,519	728
AIAN	22	95	49	556	101	823	2,588
Other	455	296	1,685	3,453	725	6,614	2,262
Ethnicity							
Hispanic	518	961	5,892	2,718	1,591	11,680	17,958
Non-Hispanic	56,310	50,505	50,669	82,016	20,285	259,785	34,002
Information Extraction Tasks	No. of Predicted Classes/No. of Total Classes						
Site							69/70
Subsite							314/327
Histology							557/644
Behavior							4/4
Laterality							7/7

Race characteristics: American Indian or Alaska Native (AIAN) and Native Hawaiian or Other Pacific Islander (NHPI); the variable *Other* contains patients who indicated other races, multiracial, or unknown.
In-Distribution (ID) registries: Kentucky (KY), Louisiana (LA), New Jersey (NJ), Washington (WA), and Utah (UT)
Out-of-Distribution (OOD) registry: New Mexico (NM)

available information (e.g., cancer pathology reports, images, laboratory results) to manually annotate the ground-truth labels of important cancer information, including primary cancer site, subsite, laterality, histology, and behavior. This annotation style, however, can result in a mismatch between the ground-truth label and the content of individual reports for CTCs with multiple reports. To decrease the number of mismatches, we removed all CTCs that were annotated as *Distant site(s)/node(s) involved*, which indicate metastasizing cancer based on the SEER summary stage classification scheme. This resulted in 2,201,126 individual reports with 1,137,993 distinct CTCs. To reduce the computational burden of our study, we limited our ID experiments to a randomly selected subset of 25% of the total CTCs, or 271,465 CTCs with 523,411 cancer pathology reports, and we removed all reports associated with 59 CTCs because they were missing demographic information, which indicates low-quality electronic health records. We then split this final dataset into a training set (189,986 CTCs/366,112 reports), validation set (40,734/79,088), and testing set (40,745/78,211) while ensuring that all reports associated with a distinct CTC are only present in one of the three splits. For the OOD experiments, we removed reports associated with 11 CTCs because they were missing demographic information; this resulted in an OOD testing dataset consisting of 106,497 cancer pathology reports.

Modern Attention Mechanism Model

Attention mechanisms are a core component in the architecture of modern clinical text classification models and usually follow the implementation proposed by Vaswani et al. [142]. Modern attention mechanisms utilize three input matrices—keys (K), values (V), and queries (Q)—to dynamically identify important parts in the latent document representation H. $K \in \mathbb{R}^{N \times d_h}$ (Equation A.1) and $V \in \mathbb{R}^{N \times d_h}$ (Equation A.2) are linear transformations of H and encapsulate important features of the text input sequence using the weights $W_K \in \mathbb{R}^{d_h \times d_h}$ and $W_V \in \mathbb{R}^{d_h \times d_h}$, biases $b_K \in \mathbb{R}^{d_h}$ and $b_V \in \mathbb{R}^{d_h}$, and a non-linear exponential linear unit (ELU) activation function. In contrast, Q depends on the purpose of the utilized attention mechanism.

If the attention mechanism (e.g., self-attention in transformers) is used to encode the contextual information of an embedded document X_E, then $Q_{self} \in \mathbb{R}^{N \times d_h}$ is another linear transformation (Equation A.3) of H using weights $W_Q \in \mathbb{R}^{d_h \times d_h}$, biases $b_Q \in \mathbb{R}^{d_h}$, and a non-linear ELU activation function. However, if the attention mechanism is used for finding the most important parts in a document context-vector representation C, then Q becomes a randomly initialized and trainable matrix that contains either task-specific (i.e., target-attention) $Q_{target} \in \mathbb{R}^{1 \times d_h}$ or label-specific (i.e., label-attention) $Q_{label} \in \mathbb{R}^{|L| \times d_h}$ reference information. In both cases, however, K and Q are used to discover connections between the words in a clinical document and the query-related reference information. The output of the comparison between K and Q is then used to assign the relative importance to each word or element in V (Equation A.4).

$$K = ELU(H \cdot W_K^\top + b_K) \tag{A.1}$$

$$V = ELU(H \cdot W_V^\top + b_V) \tag{A.2}$$

$$Q = ELU(H \cdot W_Q^\top + b_Q) \tag{A.3}$$

$$C(Q, K, V) = softmax(\frac{Q \cdot K^\top}{\sqrt{d_h}})V \tag{A.4}$$

Baseline Text-encoder Architectures

The reader should note that conventional text-encoder architectures utilize a word-based vocabulary and require manual pretraining of the word embedding layer using the documents specific to each dataset, whereas transformer-based text-encoder architectures* utilize byte-pair encoded vocabularies and typically come with their own pretrained subword embedding layer. For both datasets, the maximum document length was set to 3,000 tokens for models that use word-based vocabularies and to 4,096 tokens for architectures that use subword-based vocabularies.

*For our experiments, we utilized the transformer library HuggingFace: https://huggingface.co/.

Convolutional Neural Network The first text-encoder architecture described here is a shallow convolutional neural network (CNN) based on the work of Kim et al. [74]. The CNN encodes each embedded document X_E via three parallel 1D convolution layers with kernel sizes of three, four, and five words to allow it to learn multiple n-grams. The three outputs are then concatenated and activated with a rectified linear unit function to generate the latent document representation $H \in \mathbb{R}^{N \times 3d_h}$.

Bi-Directional Gated Recurrent Unit Neural Network The second text-encoder architecture is the two-layered bi-directional gated recurrent unit neural network (Bi-GRU) [31]. The Bi-GRU takes an embedded document X_E as the input and sequentially learns a latent document matrix $H \in \mathbb{R}^{N \times 2d_h}$.

Multitask Convolutional Neural Network The first baseline is the multitask convolutional neural network (MT-CNN) based on the works of Kim et al. [74] and Alawad et al. [7]. The MT-CNN encodes each embedded document X_e via three parallel 1D convolution layers with kernel sizes of three, four, and five words to allow it to learn multiple n-grams. The three outputs are then concatenated to generate $H \in \mathbb{R}^{N \times 3d_h}$ and passed to a max-pooling layer that generates a fixed-length document context vector representation $C \in \mathbb{R}^{1 \times 3d_h}$. This context vector representation is then passed to five parallel linear classification layers that simultaneously generate predictions for the five information extraction tasks: site, subsite, laterality, histology, and behavior.

Multitask Hierarchical Self-Attention Network The second baseline is the multitask hierarchical self-attention network (MT-HiSAN) based on the work of Gao et al. [50] and is to our knowledge the current state-of-the-art model for information extraction tasks in cancer pathology reports. The MT-HiSAN deploys a two-level hierarchical architecture that consists of multiple self-attention and target-attention

mechanisms to learn a document's context vector representation. For the MT-HiSAN, each cancer pathology is split into individual lines consisting of 15 words that are then processed by the lower hierarchy to learn a vector representation for each line $H_{line} \in \mathbb{R}^{1\times d_h}$. All line vector representations H_{line} are then passed to the upper hierarchy, which consolidates the learned line-level information into a final document context vector representation $C \in \mathbb{R}^{1\times d_h}$. This context vector representation is then passed to five parallel linear classification layers to simultaneously generate predictions for the five information extraction tasks: primary cancer site, subsite, laterality, histology, and behavior.

Transformer-Based Clinical Longformer The final baseline is the base version of the Clinical Longformer (CLF-BS) [84] implemented via the Hugging Face open-source framework for deep learning and transformers. The base version of the CLF embeds and encodes the word embedding matrix X_E to learn a latent document representation $H \in \mathbb{R}^{N\times d_h}$ with hidden dimension $d_h = 768$. Unlike the CLF-DPLA, the base version of the CLF uses the latent vector representation of the first special token $H_0 \in \mathbb{R}^{1\times d_h}$ of the text sequence as the final document context vector representation. H_0 is passed directly into the classification layer without any further feature extraction.

Quantitative Evaluation

We evaluated the performance of all models using quantitative metrics established in the multiclass text classification literature and measured the performance accuracy (Equation B.2) and macro-averaged F1-score (Equation B.3) [76, 49, 97]. The macro-averaged F1-score measures the averaged unweighted performance of all per-class F1-scores (Equation B.4) to quantify a model's ability to classify minority classes. We use the widely accepted normal approximation [115] to compute 95% confidence intervals:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{A.5}$$

$$\text{Macro-F1-Score} = \frac{1}{C} \sum_{c=i}^{C} F1\text{-}Score_c \tag{A.6}$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{A.7}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

$$\text{F1-Score}_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c},$$

where C is the total number of possible classes in the task-specific label space, c is the i^{th} class, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Model Training and Hyperparameter Optimization

Multi-Class Experiments: SEER Electronic Pathology Reports We fine-tuned our CLF-models until convergence from the pretrained weights of the base version of the CLF [84]. We optimized the model weights using the AdamW optimizer and the cross-entropy loss. The training process was parallelized across four compute nodes, each of which boasted four NVIDIA 32 GB Tesla V100 GPUs. We further deployed a linear learning rate scheduler to stabilize the fine-tuning process for the first five epochs and applied early stopping after three epochs to avoid overfitting. In contrast, the MT-HiSAN and MT-CNN were trained on a single compute node equipped with four NVIDIA 16 GB Tesla V100 GPUs using the Adam optimizer and an early stopping mechanism of five epochs. We enabled automatic mixed precision to accelerate the training process across all model architectures.

We optimized the hyperparameters of all evaluated models on the validation split for the respective dataset by using a hill-climbing strategy. This strategy reduces the computational burden by starting from a known well-performing set of

hyperparameter values and then optimizes the value of one hyperparameter at a time. For the multi-class experiments on the electronic pathology reports (see Table A.3), the final hyperparameter value was selected based on the model with the best macro- and micro-averaged F1-scores averaged across all five information extraction tasks (i.e., site, subsite, laterality, histology, and behavior). For the MT-CNN and MT-HiSAN, we followed the hyperparameter values with slight deviations proposed in the work of Gao et al. [49], who extensively tested both baselines on our SEER cancer pathology reports (see Table A.3).

Multi-Label Experiments: MIMIC-III Hospital Discharge Summary Notes We trained all tested models—CNN, Bi-GRU, and CLF—until convergence; the CLF models started from the pretrained weights of CLF's base version [84]. We selected the Adam optimizer for the CNN and Bi-GRU and the AdamW optimizer for the CLF and computed the loss using the binary cross-entropy objective function. The training process was accelerated using automatic mixed precision and parallelized with an NVIDIA DGX-1 system equipped with four NVIDIA 32 GB Tesla V100 GPUs. We further deployed a linear learning rate scheduler to stabilize the fine-tuning process for the first five epochs. We applied early stopping after five epochs for the CLF and after ten epochs for the CNN and Bi-GRU models. We tuned the hyperparameters of all evaluated models following the same strategy as described in section A.1, but selected the final value for a given hyperparameter based on the largest mean micro- and macro-averaged F1-scores. The selected hyperparameters are shown in Table A.4.

A.2 Discussion

Multi-Label Classification – Previous State-of-the-Art Models

This section provides an overview of current and previous state-of-the-art models for the MIMIC-III-Full and MIMIC-III-50 multi-label clinical text classification tasks

Table A.3: Hyperparameter optimization for the multitask convolutional neural network (MT-CNN), the multitask hierarchical self-attention network (MT-HiSAN), and the Clinical Longformer using the base-version (CLF-BS) and the deformable phrase-level attention mechanism (CLF-DPLA). Selected values are bold.

MT-CNN	
Learning Rate	5e-5, **1e-4**, 2e-4, 5e-4
Dropout %	0, 0.15, 0.3, **0.5**
Filter Size	100, **300**, 500, 1000
Batch Size	32, 64, **128**, 256
MT-HiSAN	
Learning Rate	5e-5, **1e-4**, 2e-4, 5e-4
Attention Dropout %	0, **0.1**, 0.15, 0.25, 0.5
Attention Size	**400**, 512, 768, 1024
Attention Heads	4, **8**, 16
Batch Size	32, 64, **128**, 256
CLF-BS	
Learning Rate	**1e-5**, 2e-5, 5e-5
Dropout %	0, 0.15, 0.3, **0.5**
Batch Size	2, **4**, 6
CLF-DPLA	
Learning Rate	**1e-5**, 2e-5, 5e-5
Dropout %	0, 0.15, 0.3,**0.5**
Attention Dropout %	0, **0.1**, 0.15, 0.25, 0.5
Maximum Context Range	5, **10**, 15, 20, 25
Batch Size	2, **4**, 6

Table A.4: Selection of hyperparameter values for the MIMIC-III models, including the convolutional neural network (CNN), the bi-directional gated unit recurrent neural network (Bi-GRU), and the Clinical Longformer (CLF). We indicate the values selected for the label-attention mechanism as L and for the deformable phrase-level attention mechanism as P.

CNN	
Filter Size	$100^{L,P}$, 300, 500, 1,000
Word Embedding Dimension	100^{L}, 200, 300^{P}
Batch Size	$16^{L,P}$, 32, 64, 128
Dropout %	0, 0.15, 0.3, $0.5^{L,P}$
Attention Dropout %	0^{P}, 0.1, 0.2, 0.3, 0.4, 0.5
Max Context Range	5^{P}, 10, 15, 20
Bi-GRU	
Hidden Size	128, $256^{L,P}$, 512
Word Embedding Dimension	$100^{L,P}$, 200, 300
Batch Size	$16^{L,P}$, 32, 64, 128
Dropout %	0, 0.15, 0.3, $0.5^{L,P}$
Attention Dropout %	0, 0.1^{P}, 0.2, 0.3, 0.4, 0.5
Max Context Range	5^{P}, 10, 15, 20
CLF	
Batch Size	$4^{L,P}$, 6, 8, 16
Dropout %	0^{L}, 0.15, 0.3, 0.5^{P}
Learning Rate	1e-5, 2e-5^{P}, 5e-5^{L}
Attention Dropout %	0, 0.1, 0.2, 0.3^{P}, 0.4, 0.5
Max Context Range	5, 10^{P}, 15, 20

(see Tables A.5 and A.6). We present these models as a supplement and discuss the performance differences versus our proposed models that deploy the deformable phrase-level attention mechanism.

Analysis of Fixed vs. Deformable Phrase-Level Attention

Our method proposes a phrase-level attention mechanism that utilizes masked self-attention to selectively include relevant contextual information for each word in the text sequence by using attention masks. This attention mask may be generated by dynamically determining the context size for each word or by setting the context size R to a predefined constant value. Previous work on various non-medical datasets suggests that deploying a variable context size is better than a fixed context size, which is also known as local attention [92]. To better understand our method in the clinical text classification space, we performed experiments on the SEER electronic pathology reports for the three important oncological information extraction tasks—site, subsite, and histology—and for the automated medical encoding of hospital discharge summary notes from the smaller MIMIC-III-50 subset.

For MIMIC-III-50 (see Table A.7), the deformable phrase-level attention mechanism performed marginally better for the CNN and Bi-GRU but significantly better for the CLF than the models extended with the fixed phrase-level attention mechanism.

For the selected oncological tasks (see Table A.8), the fixed phrase-level attention mechanism performs slightly better than the deformable phrase-level attention mechanism; however, the performance difference is marginal.

Table A.5: Current and previous state-of-the-art models for the MIMIC-III-Full clinical text classification task.

Model	Attention Mechanism	AUC		F1		P@k	
		Macro	Micro	Macro	Micro	8	15
CAML [99]	Label	0.895	0.986	0.088	0.539	0.709	0.561
DR-CAML [99]	Label	0.897	0.985	0.086	0.529	0.690	0.548
LAAT [143]	Label	0.919	0.988	0.990	0.575	0.738	0.591
JointLAAT [143]	Label	0.921	0.988	**0.107**	0.575	0.735	0.590
EffectiveCAN [88]	Label	0.921	**0.989**	0.105	0.581	0.755	0.604
PLM-ICD* [61, 39]	Label	**0.926**	**0.989**	0.104	**0.598**	**0.771**	**0.613**

* Original authors [61] only provided results for MIMIC-III-Full; for completion showing reproduced results of MIMIC-III-50 [39].

Table A.6: Current and previous state-of-the-art models for the MIMIC-III-50 clinical text classification task.

Model	Attention Mechanism	AUC		F1		P@k
		Macro	Micro	Macro	Micro	5
CAML [99]	Label	0.875	0.909	0.532	0.614	0.609
DR-CAML [99]	Label	0.884	0.916	0.576	0.633	0.618
LAAT [143]	Label	0.925	0.946	0.666	0.715	0.675
JointLAAT [143]	Label	0.925	0.946	0.661	0.716	0.671
EffectiveCAN [88]	Label	0.920	0.945	0.668	0.717	0.664
HiLAT + ClinicalplusXLNet [87]	Label	**0.927**	**0.950**	**0.690**	**0.735**	**0.681**
PLM-ICD* [61, 39]	Label	0.917	0.938	0.663	0.705	0.657

* Original authors [61] only provided results for MIMIC-III-Full; for completion showing reproduced results of MIMIC-III-50 [39].

Table A.7: Preliminary results on the fixed (FPLA) vs. deformable (DPLA) context size for the proposed phrase-level attention mechanism on MIMIC-III-50. We tuned the context size R for each text-encoder architecture; we used the hyperparameters of the DPLA models as the starting point for tuning the fixed context size.

Model	Module	AUC		F1-Score		P@k
		Macro	Micro	Macro	Micro	5
CNN	FPLA (R=10)	0.889	0.916	0.578	0.642	0.619
	DPLA (R=5)	**0.890**	**0.917**	**0.587**	**0.648**	**0.623**
Bi-GRU	FPLA (R=10)	0.894	0.922	0.576	0.643	0.629
	DPLA (R=5)	**0.895**	**0.923**	**0.577**	**0.646**	**0.633**
CLF	FPLA (R=15)	0.905	0.928	0.603	0.665	0.647
	DPLA (R=10)	**0.910**	**0.931**	**0.634**	**0.682**	**0.650**

Convolutional neural network (CNN), bi-directional gated recurrent unit neural (Bi-GRU), Clinical Longformer (CLF)

Table A.8: Preliminary results on the fixed (FPLA) vs. deformable (DPLA) context size for the proposed phrase-level attention mechanism on important oncological information extraction tasks (site, subsite, and histology) for SEER electronic pathology reports.

Model	Module	Site		Subsite		Histology	
		Accuracy	F1-Macro	Accuracy	F1-Macro	Accuracy	F1-Macro
CLF	FPLA ($R = 5$)	**0.946**	0.736	**0.721**	**0.393**	**0.811**	0.413
	FPLA ($R = 10$)	**0.946**	**0.737**	**0.721**	0.392	0.810	**0.414**
	DPLA	**0.946**	0.734	0.720	0.388	**0.811**	0.411

Soft-Abstention Procedure

A common strategy for achieving low error rates is to use a learned rejection rule that forces the model to abstain from providing uncertain predictions during training and inference. However, in this study, we utilize soft-abstention analysis that is retrospectively applied to a model's output to discard documents with uncertain predictions according to a rejection rule while maintaining a predefined error rate. The soft-abstention procedure is described below [35]:

1. Initialize an empty array with probability thresholds ranging from 0 to 1 and a step size of 0.001 (i.e., $thresholds = [0.001, 0.002, ..., 0.999]$).

2. For each probability threshold, retain all validation documents with a softmax probability greater than or equal to the current threshold, drop the remaining validation documents, and compute the error rate for the retained validation documents using a performance accuracy score (i.e., 100% − performance accuracy = error rate).

3. Find the smallest probability threshold that results in an error rate below 3%.

4. During inference using the test dataset, use this probability threshold as the rejection rule to remove all documents with a softmax probability below the rule.

5. Compute the percentage of documents that remain in the dataset after soft abstention and measure the performance accuracy on the remaining test data.

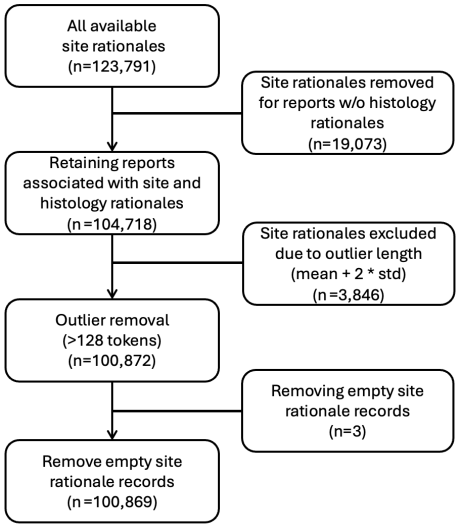
B Chapter 4 Supporting Information

B.1 Materials and Methods

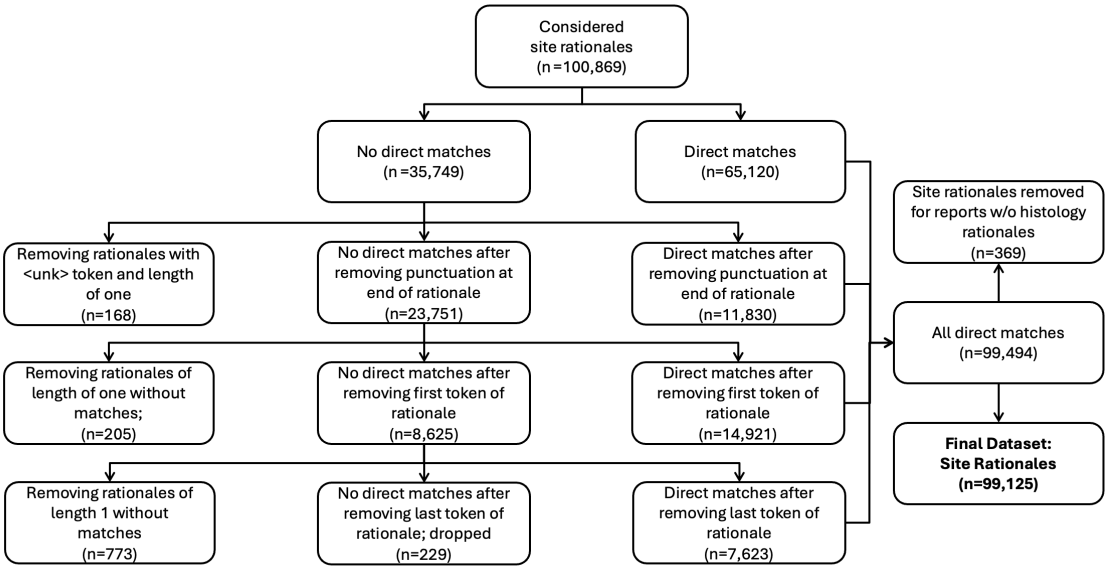
Data

This work considers human-based clinical rationales that provide plausible explanations for the primary cancer site assigned to electronic pathology reports. The included flowchart provides a detailed overview of generating the final rationale dataset (Figure B.1). Notably, although we initially considered analyzing records annotated with rationales for the primary cancer site *and* the histological type, we ultimately changed our research priority to focus on only primary cancer sites, which led us to not include rationales associated with histology. Because the experiments were already underway when we changed our strategy, we did not utilize the full set of available primary cancer site rationales, and we acknowledge a potential introduction of sampling bias when choosing our subset of primary cancer site rationales.

Generation of Human-based Clinical Rationales Participating Surveillance, Epidemiology, and End Results (SEER) registries deploy an ML-based API to auto-encode roughly 17% of all pathology reports with an accuracy above 97% [59]. To ensure high annotation quality, approximately 10% of the auto-encoded reports are manually reviewed by data oncology specialists. If a specialist disagrees with the top-ranked diagnosis suggested by the API for a given report, then they provide the correct diagnosis along with a supporting highlight from the report or a comment in the absence of the information relevant for the provided diagnosis. In our study, we focus only on highlights related to the primary cancer site that match at least one instance of the rationale in the associated report, which results in a total of 89,718 reports with 99,125 rationales. We exclude all reports annotated with comments because they may not necessarily be extractive. Notably, the exact spatial location of these highlights was lost during preprocessing because the spatial location indicated the positions



(a) Flow chart of rationale selection process.



(b) Flow chart of rationale selection process with matching instances in their associated reports.

Figure B.1: Overview of the primary cancer site rationale selection and matching process. Initially, only reports with site and histology rationales were considered. However, the histology rationales were omitted owing to a change in research priority.

(i.e., row and column position in a text file) of the first and last characters of a given highlight in the raw text file. Therefore, we considered all occurrences of a given highlight as a rationale within the report. We treat our rationales as independent samples and integrate the rationale information into the model training as additional training data, which is similar to the approaches of previous work [158, 125, 116]. Additional details on the generation of rationales can be found in [138].

We included nearly all reports with at least one rationale in the training split and reserved a small fraction for the testing split to enable us to perform model explainability analyses (Table 4.1). Additionally, models that perform the automated medical encoding of clinical documents do not have access to rationales during inference. The number of distinct rationales per document ranges from one to seven, with approximately 99% of the training reports having either one or two rationales. Each report in the test set has one rationale, allowing us to maximize the number of available rationales in the training dataset by avoiding the removal of reports with multiple rationales. For each rationale, we created binary masks that indicate whether a token is part of a rationale by matching the rationale with a substring of the report. Although the majority of the 65,120 rationales were matched directly, a total of 34,005 rationales were matched after further cleaning, and a total of 1,744 rationales were dropped because they were not matchable. See Appendix B.1 for a detailed overview of the rationale generation. Because the initial, raw set of rationales contained very long sequences, longer than 1,000 tokens with a maximum of 1,235 tokens, we assumed those greater than the average rationale length plus 2× the standard deviation—maximum rationale length is 128 tokens—were erroneous, and we excluded these from our analysis.

We used the binary masks to create complements for the rationales (i.e., complement samples include the full pathology report minus the information contained by a rationale). Similar to [26], we drop all rationale tokens from the reports to prevent the model from potentially learning positional information from the rationales. To populate the validation and testing splits, documents were randomly sampled from

the available total population of SEER cancer pathology reports (these reports do not possess any rationales). We ensured that reports associated with a distinct Cancer-Tumor-Case (CTC) were confined to a single split and that class distributions were consistent across all splits. Lastly, we randomly selected an additional set of cancer pathology reports, ensuring no overlap in CTCs and matching the same distribution as the set of rationales, to investigate whether annotations efforts should be spent on retrieving rationales or labeling new, unlabeled reports.

Table B.1 shows the distribution of records, cases, and patients by registry and data split. Because our training dataset contains documents annotated with multiple rationales, we show the breakdown in Table B.2. Each test report was only annotated with a single rationale. Table B.3 shows a breakdown of the number of rationales associated with a distinct range of rationale lengths.

Table B.1: Overview of SEER electronic pathology reports.

Registry	Train	Val	Test	Total
New Jersey	72,079	8,773	10,826	91,678
Louisiana	13,135	7,484	7,843	28,462
Utah	2,038	3,128	3,343	8,509
Reports	87,252	19,385	22,012	128,649
Tumor cases	51,136	7,916	10,351	69,403
Patients	49,616	7,910	10,341	67,867

Table B.2: Frequency distribution of human-created rationales for SEER cancer pathology reports contained in the training dataset.

Number of Rationales per Document	Number of Documents	Total Number of Rationales
1	78,801	78,801
2	7,608	15,216
3	736	2,208
4	92	368
5-7	15	86
Total	87,252	96,679

Table B.3: Frequency distribution of rationale length in word-level tokens of human-created rationales associated with SEER cancer pathology reports. The class-wise proportion of the training rationales for the 10 common primary cancer sites shows a breakdown in rationale length. The maximum length of a rationale is 128 words.

Rationale Length Range in Word Tokens	Split		Class-Wise Training Rationale Proportions in %									
	Train	Test	C16	C18	C25	C42	C44	C50	C54	C61	C64	C73
1-2	16,220	452	23.3	16.6	18.2	24.5	12.9	13.4	26.5	50.4	15.3	21.6
3-5	17,322	429	16.5	15.1	16.3	17.8	28.3	15.0	21.4	13.8	13.8	24.0
6-10	16,812	433	25.1	23.8	19.3	13.7	18.9	14.3	16.8	12.0	23.2	15.6
11-15	16,503	452	14.6	15.3	18.6	14.0	14.5	18.5	11.7	8.5	21.4	14.0
16-25	14,185	359	9.1	10.7	13.6	14.6	10.0	18.4	9.9	4.8	12.3	11.2
26+	15,637	321	11.3	18.6	14.1	15.4	14.3	20.5	13.7	10.5	14.0	13.5
Total	96,679	2,446										

Ten common primary cancer sites: stomach (C16), colon (C18), pancreas (C25), hematopoietic and reticuloendothelial systems (C42), skin (C44), breast (C50), corpus uteri (C54), prostate gland (C61), kidney (C64), and thyroid gland (C73).

Model Architectures

Here, we briefly describe the evaluated model architectures. We selected the clinical longformer (CLF) with its baseline weights (CLF-BS) as the text-encoder architecture for all models because the CLF has enabled models to achieve exceptional performance on clinical text classification tasks [84]. Our specific implementation is from the HuggingFace transformer library. The CLF utilizes a byte-pair encoded vocabulary to tokenize the incoming text sequences. A pretrained subword embedding layer is used to transform each subword token into a vector representation with a hidden dimension of $d_h = 768$. The CLF-BS is the baseline model for all experiments. The second evaluated model utilizes a recently proposed multiple attention mechanism to learn lexical and contextual information from the input sequence relevant to our classification task [94].

Given an input text sequence $X = [x_1, x_2, \ldots, x_N]$ with N tokens, each token is mapped to its corresponding embedding vector representation with size d_e, creating a token-embedding matrix $X_E \in \mathbb{R}^{N \times d_e}$. X_E is then processed by a text-encoder architecture (i.e., CLF) to learn a latent document matrix $H \in \mathbb{R}^{N \times d_h}$, where d_h is the

hidden dimension of the underlying architecture. Depending on the evaluated model architecture—CLF-BS or CLF with a deformable phrase-level attention mechanism (CLF-DPLA)—the latent document matrix H is either used directly as the final document context vector representation $C \in \mathbb{R}^{1\times d_h}$ or further processed. For the CLF-BS, we follow standard practice with transformers and used the hidden states $H_0 \in \mathbb{R}^{1\times d_h}$ of the first special token <s> as C. In contrast, the CLF-DPLA deploys a complex attention mechanism to further encode H to enrich C with lexical word-level and contextual phrase-level information relevant to the classification task. C is then passed to a linear classification layer (Equation B.1) with projection weights $W \in \mathbb{R}^{|L|\times d_h}$ and bias $B \in \mathbb{R}^{|L|\times 1}$. $|L|$ is the number of classes for a given task. The output of the classification layer is then passed to a softmax activation function, thereby creating class-wise pseudo-probabilities. The class with the maximum probability is selected as the prediction $\hat{y}$ for a given text sequence. We refer the reader to [94] for more details. All text sequences were tokenized by utilizing the byte-pair encoded vocabulary associated with the CLF, and the maximum sequence length was 4,096 tokens.

$$\hat{y}(C, W, B) = \arg\max_{y} softmax(C \cdot W^{\top} + B) \tag{B.1}$$

B.2 Model Training

Starting from the pretrained weights of the CLF-BS, we fine-tune our CLF models to convergence [84]. The model weights are optimized by using the AdamW optimizer and the cross-entropy objective loss function. Training is accelerated by using automatic mixed precision and is parallelized across two compute nodes of the Frontier supercomputer via the CITADEL security framework. Because SEER electronic pathology reports contain protected health information, we utilized the CITADEL security framework associated with the Scalable Protected Infrastructure provided by the National Center for Computational Sciences and the Oak Ridge Leadership

Computing Facility. Each Frontier compute node is equipped with four AMD MI250X GPUs (with two Graphics Compute Dies in each GPU), which are equal to eight 64 GB memory GPUs. To stabilize the fine-tuning process, we employed a linear learning rate scheduler for the first five epochs and implemented early stopping after three epochs to prevent overfitting. We utilized the hyperparameters identified in [94] and set the batch size to eight.

B.3 Model Evaluation

We evaluate the performance of all models using quantitative metrics established in the multiclass text classification literature and measured the performance accuracy (Equation B.2) and macro-averaged F1-score (Equation B.3) [76, 49, 97]. We use the widely accepted normal approximation [115] to compute 95% confidence intervals.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{B.2}$$

$$\text{Macro-F1-Score} = \frac{1}{C} \sum_{c=i}^{C} F1\text{-}Score_c \tag{B.3}$$

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{B.4}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

$$\text{F1-Score}_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c},$$

where C is the total number of possible classes in the task-specific label space, c is the i^{th} class, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

B.4 Results

Preliminary Results

During the preliminary phase of the study, we experimented with various algorithmic approaches to integrate the rationale information into the models. Primarily, we experimented with approaches that integrated rationale information as binary masks, which indicate whether or not a token is a part of a rationale. The first model utilizes a CLF-BS as the text encoder architecture to encode the input sequence and two multitask classification layers to perform token-sequence classification. The first layer performs token classification by trying to predict whether a token is part of a rationale. It does this by using the hidden states of all tokens of the sequences as input and the binary rationale masks as the ground-truth labels. The second layer utilizes the hidden states of the first special token as the document vector representation to perform its prediction for the primary cancer site. The loss from the token classification is scaled with $\alpha = 0.25$ and added to the full loss from the sequence classification. We evaluated $\alpha \in [0.25, 0.5, 0.75]$ to optimize the models' performance. The second model we investigated extends the CLF-BS with a masked multihead self-attention layer using the binary rationale mask to control the attention flow of the two self-attention heads. We optimized for the number of heads from $m_{heads} = [2, 4, 8]$. Conceptually, this model is designed to optimize its parameters to focus on information that is contextually similar to the rationales.

Table B.4 overviews the preliminary experiments performed to determine the optimal method for incorporating human-based clinical rationales into the training procedure of transformer-based clinical text classification models. We focused our experiments on the CLF because it has shown favorable performance in clinical text classification tasks [95]. The evaluated methods include data augmentation (i.e., using the rationales as additional samples in the training dataset) and changes to the underlying algorithm following proposed models in the literature. For the algorithmic approach, we explored the use of binary masks to indicate whether a token is part

of a rationale as an additional input to a multitask, learning-based token-sequence classification model. These masks functioned as the ground-truth labels for token classification and as an attention mask used in a masked multihead self-attention head. The preliminary experiments show that the addition of rationales to the training dataset outperformed the algorithmic implementations, and this is why we selected the rationales as the primary approach to incorporating human expert–derived input in ML model training.

Table B.4: Preliminary results for incorporating human-based rationales in the CLF.

Clinical Longformer	Training Data Input	F1-Macro	Accuracy
Baseline	Reports only	0.547	0.881
+ Deformable Phrase-Level Attention	Reports only	0.582	0.898
Baseline	Reports and Rationales	0.562	0.885
+ Deformable Phrase-Level Attention	Reports and Rationales	**0.592**	**0.902**
+ Token-Sequence-Classification	Reports and Rationale Masks	0.525	0.881
+ Masked Multi-Head Self-Attention	Reports and Rationale Masks	0.573	0.897
Dataset sizes: $n_{train} = 90,660$, $n_{train,rationales} = 100,204$, $n_{val} = 19390$, and $n_{test} = 19,522$)			

Results: Performance

Table B.5: Improvements in accuracy and F1-macro scores with the addition of m randomly selected, *sufficient*, human-generated rationales associated with 10 common primary cancer sites for the CLF-BS and CLF-DPLA models when classifying primary cancer sites from SEER cancer pathology reports. The ↑ and ↓ indicate increased or decreased performance, respectively, compared to the model trained with only reports. Scores marked with – have equal performance.

Model	Number of Added Rationales	Metric	Sufficiency Threshold		
			> 0	> 0.2	> 0.817
CLF-BS	100	Accuracy	0.884↓ (0.880, 0.888)	0.884↓ (0.880, 0.888)	0.885 – (0.880, 0.889)
		F1-Macro	0.555↓ (0.549, 0.562)	0.565↓ (0.558, 0.572)	0.564↓ (0.557, 0.571)
	500	Accuracy	0.886↑ (0.881, 0.890)	0.882↓ (0.878, 0.887)	0.884↓ (0.880, 0.889)
		F1-Macro	0.568↑ (0.561, 0.574)	0.563↓ (0.556, 0.569)	0.565↓ (0.558, 0.571)
	1,000	Accuracy	0.884↓ (0.880, 0.889)	0.885 – (0.880, 0.889)	0.884↓ (0.884, 0.884)
		F1-Macro	0.570↑ (0.563, 0.577)	0.565↓ (0.558, 0.571)	0.570↑ (0.563, 0.577)
CLF-DPLA	100	Accuracy	0.898↓ (0.894, 0.902)	0.899↓ (0.895, 0.903)	0.900 – (0.896,v0.904)
		F1-Macro	0.583↓ (0.576, 0.590)	0.582↓ (0.575, 0.589)	0.584↓ (0.577, 0.591)
	500	Accuracy	0.900 – (0.896, 0.904)	0.899↓ (0.895, 0.903)	0.901↑ (0.897, 0.905)
		F1-Macro	0.589↓ (0.583, 0.595)	0.596↓ (0.589, 0.602)	0.580↓ (0.574, 0.587)
	1,000	Accuracy	0.898↓ (0.894, 0.902)	0.899↓ (0.895, 0.903)	0.897↓ (0.893, 0.901)
		F1-Macro	0.590↓ (0.583, 0.596)	0.580↓ (0.574, 0.587)	0.580↓ (0.573, 0.586)

Common primary cancer sites: stomach (C16), colon (C18), pancreas (C25), hematopoietic and reticuloendothelial systems (C42), skin (C44), breast (C50), corpus uteri (C54), prostate gland (C61), kidney (C64), and thyroid gland (C73). Sufficiency thresholds: >0 includes all rationales, >0.2 is the best achievable performance improvement, and >0.817 represents the 90% percentile of rationales with false predictions for falsely predicted reports. The 95% confidence interval is computed via normal approximation [115]. [a]Utilized all 973 rationales associated with kidney cancer (C64) after removing all non-sufficient rationales.

Results: Explainability

Average Token-Level Rationale Coverage: Prediction

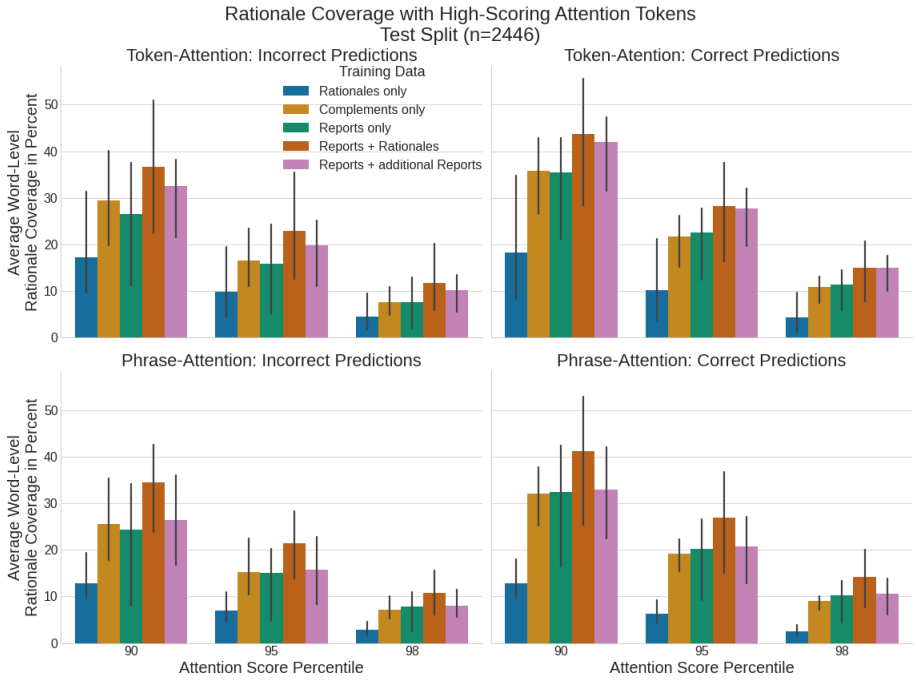


Figure B.2: Average word-level rationale coverage broken down by training type and prediction type (correct vs. incorrect) based on attention scores generated by the token- and phrase-level target attention mechanisms of the DPLAs implemented in conjunction with the CLF for different attention score percentiles.

Average Token-Level Rationale Coverage: Rationale Length

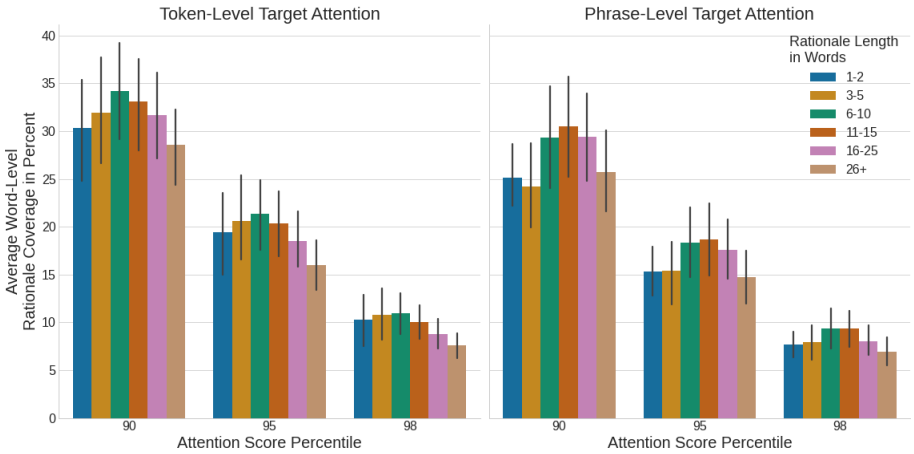
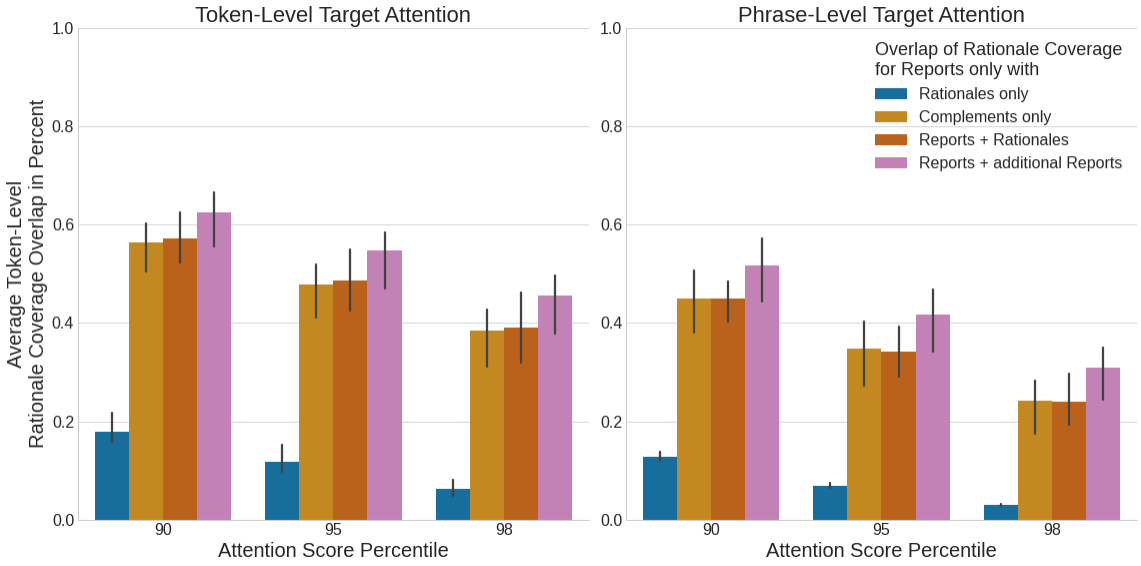


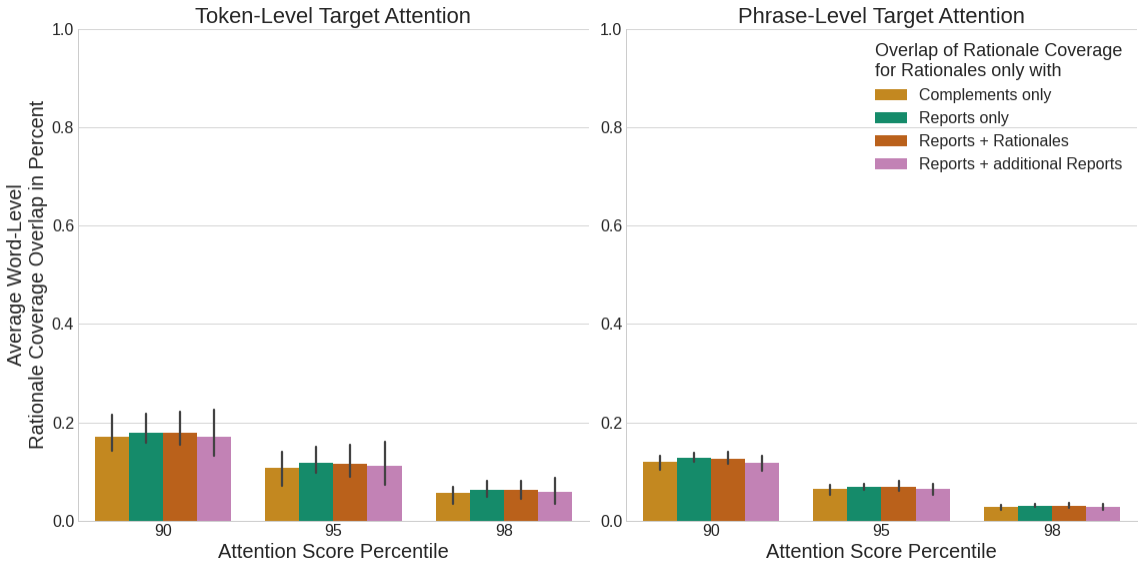
Figure B.3: Average word-level rationale coverage broken down by rationale length based on attention scores generated by the token- and phrase-level target attention mechanisms of the DPLAs implemented in conjunction with the CLF for different attention score percentiles.

Average Token-Level Rationale Coverage: Overlap

The following figures show how models trained on the various regimens highlight up to 60% of the same tokens associated with the rationales. Each figure represents the view from one training regimen compared with the remaining four training regimens.

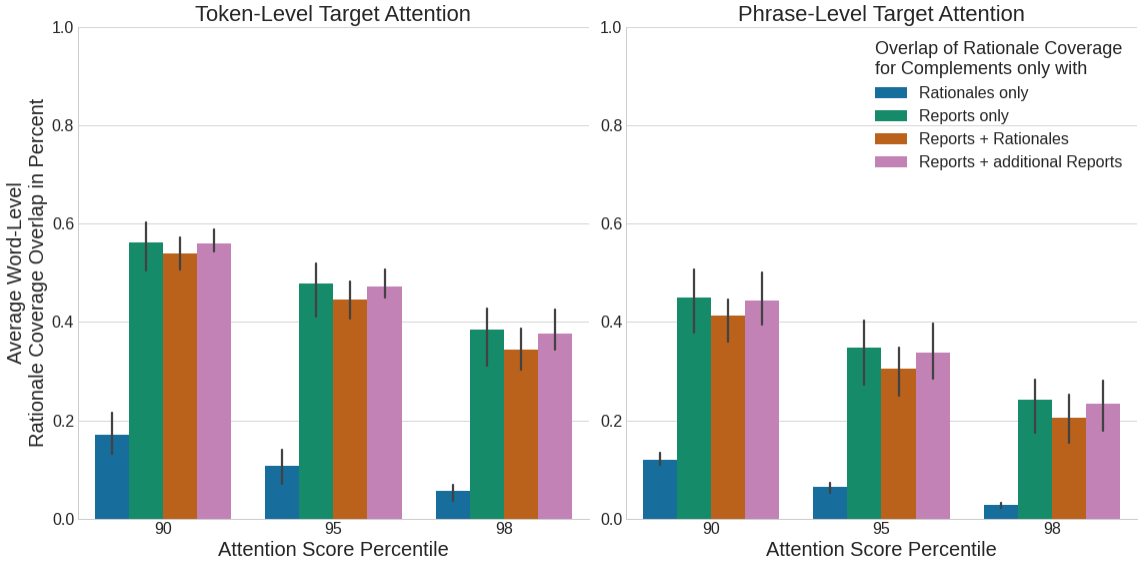


(a) Overlap of rationale coverage for models trained on reports only.

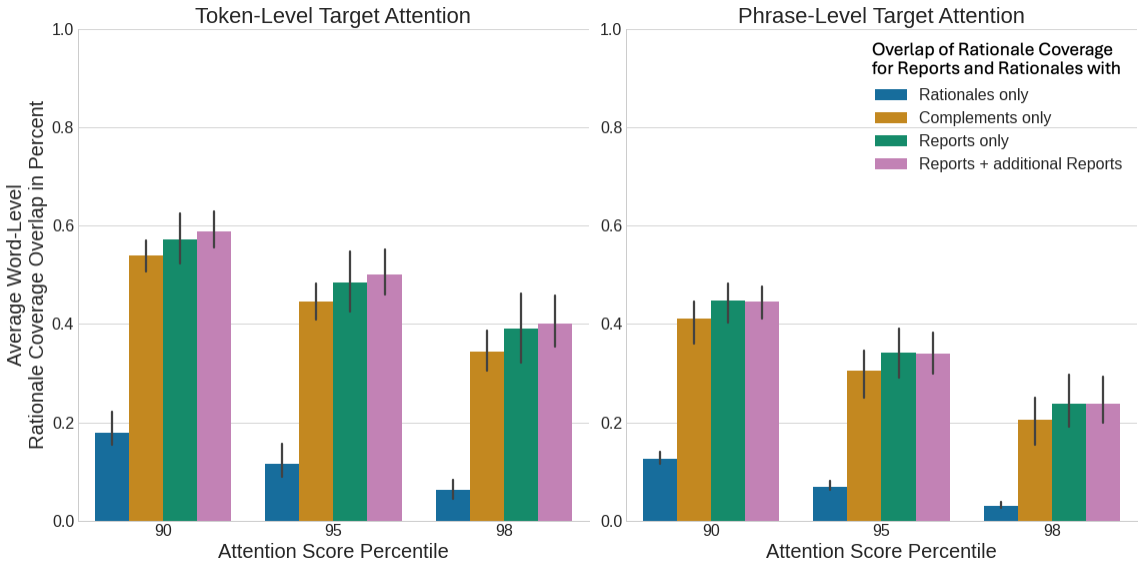


(b) Overlap of rationale coverage for models trained on rationales only.

Figure B.4: Average word-level rationale coverage overlap on test documents associated with clinical rationales for models trained on reports only. Rationale coverage computed based on the attention scores generated by the token- and phrase-level target attention mechanisms of the DPLAs implemented in conjunction with the CLF for different attention score percentiles.

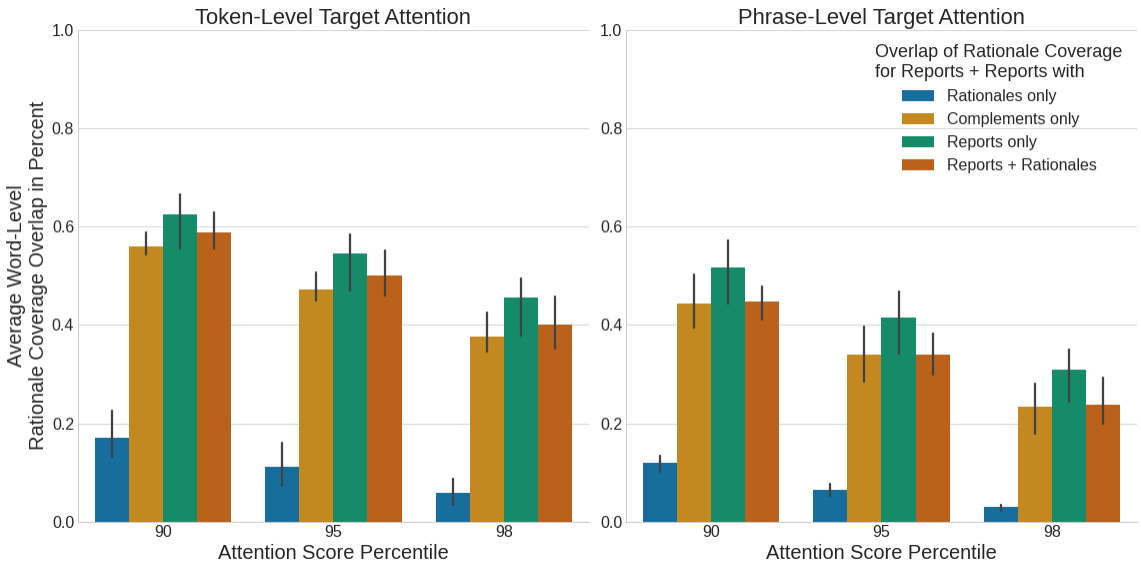


(c) Overlap of rationale coverage for models trained on complements only.



(d) Overlap of rationale coverage for models trained on reports supplemented with human-based clinical rationales.

Figure B.4 (cont.)**:** Overlap of the average token-level rationale coverage for test documents associated with clinical rationales across the five training regimens; complements only and reports supplemented with human-based clinical rationales.



(e) Overlap of rationale coverage for models trained on reports supplemented with additional reports.

Figure B.4 (cont.)**:** Overlap of the average token-level rationale coverage for test documents associated with clinical rationales across the five training regimens; reports supplemented with additional electronic pathology reports.

Vita

Christoph Simon Metzner, born in Traunstein, Germany, began his academic journey with a bachelor's degree in Wood Technology. His interests evolved during a joint-degree program at the University of Applied Sciences Salzburg, Austria, and the University of Tennessee, Knoxville, USA, where he discovered a passion for developing software programs and data science. This newfound interest led him to pursue a Ph.D. in Data Science and Engineering, focusing on advancing clinical natural language processing and the automated encoding of electronic pathology reports.