Computer Science

Efficient Transfer Learning with Sequential and Multi-Modal Approaches for Electronic Health Records

Yogesh Kumar

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 197/2024

Efficient Transfer Learning with Sequential and Multi-Modal Approaches for Electronic Health Records

Yogesh Kumar

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall U154 U1 VALMET of the school on 25 October 2025 at Aalto University.

Aalto University
School of Science
Computer Science
Machine Learning for Health (ML4H)

Supervising professor
Dr. Pekka Marttinen, Aalto University, Finland

Thesis advisor
Dr. Pekka Marttinen, Aalto University, Finland

Preliminary examiners
Dr. Mehul Motani, National University of Singapore
Dr. Catherine Jutzeler, ETH Zurich, Switzerland

Opponent
Dr. Line Clemmensen, Technical University of Denmark (DTU)

Aalto University publication series
DOCTORAL THESES 197/2024

ISBN 978-952-64-2030-1 (printed)
ISBN 978-952-64-2031-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-2031-8

Unigrafia Oy
Helsinki 2024

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Author
Yogesh Kumar

Name of the doctoral thesis
Efficient Transfer Learning with Sequential and Multi-Modal Approaches for Electronic Health Records

Publisher School of Science

Unit Computer Science

Series Aalto University publication series DOCTORAL THESES 197/2024

Field of research Machine Learning

Manuscript submitted 14 May 2024 **Date of the defence** 25 October 2024

Permission for public defence granted (date) 30 August 2024 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

The digital transformation in healthcare has dramatically increased data availability, yet the potential for data-driven insights is frequently constrained by the quality of data. Securing high-quality data is particularly challenging in fields like healthcare, where expert involvement is crucial for gathering, annotating, and ensuring data quality. This thesis applies deep learning to Electronic Health Records (EHR) to enhance predictive accuracy and operational efficiency. Deep learning models are particularly adept at capturing complex and non-linear relationships present in EHR data, but they require extensive training datasets to be effective. This study explores and develops ways to employ transfer learning to effectively mitigate these data constraints.

The thesis tackles four key research questions aimed at improving healthcare outcomes using EHR data. It begins by enhancing prediction accuracy for healthcare utilization through an RNN model with multi-headed attention, which significantly outperforms traditional count-based models and shows robust time generalization. The study then introduces SANSformer, a custom-built, attention-free sequential model optimized for EHR specifics. This model excels in predicting healthcare demand, particularly in diverse patient subgroups, while managing limited data scenarios via transfer learning. Thirdly, the thesis explores the enhancement of neural network similarity metrics in assessing functional similarities, particularly in the context of transfer learning and model performance. It introduces a covariate adjustment to correct traditional metrics, which are often misled by input data structures, ensuring they reflect true functional similarities. Lastly, it explores the integration of expert annotations into the medical CLIP model, eCLIP, which utilizes radiologist eye-gaze heatmaps to substantially improve the quality of embeddings and sample efficiency in multi-modal medical imaging.

The findings from this thesis highlight the significant potential of deep learning to enhance prediction of healthcare outcomes by addressing the unique challenges of EHR data. The research adapts sophisticated deep learning models to meet the complex demands of EHR data and introduces novel techniques like covariate adjustment for similarity metrics and integrating expert annotations to set a foundation for further advancements in healthcare analytics.

Keywords Deep Learning, Transfer Learning, Healthcare

ISBN (printed) 978-952-64-2030-1 **ISBN (pdf)** 978-952-64-2031-8

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2024

Pages 156 **urn** http://urn.fi/URN:ISBN:978-952-64-2031-8

Abstract

The digital transformation in healthcare has dramatically increased data availability, yet the potential for data-driven insights is frequently constrained by the quality of data. Securing high-quality data is particularly challenging in fields like healthcare, where expert involvement is crucial for gathering, annotating, and ensuring data quality. This thesis applies deep learning to Electronic Health Records (EHR) to enhance predictive accuracy and operational efficiency. Deep learning models are particularly adept at capturing complex and non-linear relationships present in EHR data, but they require extensive training datasets to be effective. This study explores and develops ways to employ transfer learning to effectively mitigate these data constraints.

The thesis tackles four key research questions aimed at improving healthcare outcomes using EHR data. It begins by enhancing prediction accuracy for healthcare utilization through an RNN model with multi-headed attention, which significantly outperforms traditional count-based models and shows robust time generalization. The study then introduces SANSformer, a custom-built, attention-free sequential model optimized for EHR specifics. This model excels in predicting healthcare demand, particularly in diverse patient subgroups, while managing limited data scenarios via transfer learning. Thirdly, the thesis explores the enhancement of neural network similarity metrics in assessing functional similarities, particularly in the context of transfer learning and model performance. It introduces a covariate adjustment to correct traditional metrics, which are often misled by input data structures, ensuring they reflect true functional similarities. Lastly, it explores the integration of expert annotations into the medical CLIP model, eCLIP, which utilizes radiologist eye-gaze heatmaps to substantially improve the quality of embeddings and sample efficiency in multi-modal medical imaging.

The findings from this thesis highlight the significant potential of deep learning to enhance prediction of healthcare outcomes by addressing the unique challenges of EHR data. The research adapts sophisticated deep learning models to meet the complex demands of EHR data and introduces novel techniques like covariate adjustment for similarity metrics and integrating expert annotations to set a foundation for further advancements in healthcare analytics.

Preface

"Between the wish and the
thing, the world lies waiting."

– Cormac McCarthy

I had imagined that the PhD journey would resemble a Western – a lone hero embarking on a lengthy and trecherous quest to get a doctorate. In reality, while the journey was indeed long and challenging, I was far from alone. Which is why am particularly excited to write this section of my thesis, where I can finally express my gratitude to those who supported me along the way.

My advisor, Pekka Marttinen, is one of the kindest people I have worked with. Over the years, I've learned so much from him – technical knowledge, of course, but also invaluable life skills, like seeing the flip side of tough situations and learning from them. Whenever we received scathing reviews on a manuscript, there would always be a message from Pekka in my inbox, carefully breaking down the feedback into manageable, objective chunks. It never failed to calm me down. Thank you for all the support, Pekka. I hope I wasn't too difficult to manage along the way!

I would also like to thank my collaborators – Prof. Sangita Kulathinal, Prof. Sami Kaski, Prof. Alexander Ilin, Dr. Maarit K Leinonen, Dr. Tianyu Cui, Henri Salo, Tuomo Nieminen and Kristian Vepsalainen. Working with each of you allowed me to approach problems from different angles, broadening my perspective and making me a more well-rounded researcher with a much wider understanding. Big thanks also to the wizards at the Aalto Science-IT Project, who, on more than one occasion, helped me overcome major technical hurdles in my projects.

My sincere thanks to the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI and grants 315896, 319323, 336033, 358246) and EU (H2020 grant 101016775) for supporting my PhD work. I also would like to express my gratitude to the pre-examiners of my thesis, Prof. Mehul Motani (NUS Singapore) and Prof. Catherine Jutzeler (ETH Zurich), for their valuable feedback.

I've been fortunate to have had amazing friends and peers at Aalto, especially in the Machine Learning for Health (ML4H) group, who kept me company through both the highs and lows. Thank you, Vladimir, Santosh, Caglar, Onur, Tianyu, Khaoula, Kunal, Mine, Tolou, Hans, Antti, Joel, Fanni, Tommi and all the past and present members of the ML4H group for the conversations over coffee, the shared memories, lunches, and everything in between.

Finally, I want to express my deepest gratitude to my family. My parents have been endlessly supportive, patiently waiting as I navigated the long journey of academia. To our cat Ollie, thank you for providing warmth and purrs on so many winter mornings. And to my wife, Preeti – without you, I'd probably still be a grumpy Java enterprise engineer nitpicking over pointless PRs. It turns out you knew me better than I knew myself, convincing me to take on this PhD adventure. I dedicate this thesis to you, for being my financier, mental coach, vacation planner, dance partner, hype-man, and ever-wise, Yoda from Pithoragarh.

Helsinki, September 9, 2024,

Yogesh Kumar

Contents

Preface **3**

Contents **5**

List of Publications **7**

Author's Contribution **9**

Abbreviations **11**

1. Introduction **13**

2. Background **17**
2.1 Electronic Health Records: An Overview 17
2.1.1 Structured EHR 18
2.1.2 Cohort 18
2.1.3 Challenges in EHR data 19
2.2 Data Representation for Structured EHR 19
2.2.1 Representation for Non-sequential Models 20
2.2.2 Representation for Sequential Models 21
2.3 Sequential Models for Structured EHRs 21
2.3.1 Long Short-Term Memory (LSTM) Networks . . . 22
2.3.2 Multi-headed Attention 22
2.4 Pre-training Neural Networks 23
2.4.1 Contrastive Language-Image Pre-training (CLIP) 24
2.5 Measuring Similarities between Neural Networks 25

3. Forecasting Healthcare Utilization **27**
3.1 Method 27
3.2 Key Results 29
3.3 Revisiting Research Question 1 30

4. Improving Sample Efficiency with Transfer Learning **31**
4.1 Method 32

4.2 Key Results 34
4.3 Revisiting Research Question 2 37

5. Covariate Adjusted Representation Similarity for Transfer Learning 39
5.1 Method 40
5.2 Key Results 41
5.3 Revisiting Research Question 3 43

6. Contrastive Language-Image Training with Expert Annotations 45
6.1 Method 46
6.1.1 Heatmap Processor 47
6.1.2 Mixup Augmentation 47
6.1.3 Curriculum Learning Strategy 48
6.2 Key Results 48
6.3 Revisiting Research Question 4 49

7. Conclusion 51

References 53

Publications 61

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I **Yogesh Kumar**, Henri Salo, Tuomo Nieminen, Kristian Vepsalainen, Sangita Kulathinal, Pekka Marttinen. Predicting utilization of healthcare services from individual disease trajectories using RNNs with multi-headed attention. *Machine Learning for Health (ML4H) at NeurIPS 2019 (Proceedings of Machine Learning Research)*, 116:93-111, 2020.

II **Yogesh Kumar**, Alexander Ilin, Henri Salo, Sangita Kulathinal, Maarit K Leinonen, Pekka Marttinen. Self-Supervised Forecasting in Electronic Health Records with Attention-Free Models. *IEEE Transactions on Artificial Intelligence*, vol. 5, 08:3926-3938 Aug 2024.

III Tianyu Cui, **Yogesh Kumar**, Pekka Marttinen, Samuel Kaski. Deconfounded Representation Similarity for Comparison of Neural Networks. *Advances in Neural Information Processing Systems*, 35:19138–19151, 2022.

IV **Yogesh Kumar**, Pekka Marttinen. Improving Medical Multi-modal Contrastive Learning with Expert Annotations. Accepted for publication in *18th European Conference on Computer Vision (ECCV)*, 2024.

Author's Contribution

Publication I: "Predicting utilization of healthcare services from individual disease trajectories using RNNs with multi-headed attention"

Prof. Marttinen and Prof. Kulathinal defined the initial problem statement. The author developed the methodology with guidance from both professors. Henri Salo, Tuomo Nieminen, and Kristian Vepsäläinen were responsible for data acquisition, processing, and establishing the computational setup. The author conducted all experiments and primarily wrote the manuscript, with input from all contributors.

Publication II: "Self-Supervised Forecasting in Electronic Health Records with Attention-Free Models"

Prof. Marttinen, Prof. Kulathinal and Dr. Leinonen defined the initial problem statement. The author developed the methodology with Prof. Ilin and Prof. Marttinen. Henri Salo was responsible for data management. The author conducted all preprocessing and experiments. The manuscript was primarily written by the author with input from all contributors.

Publication III: "Deconfounded Representation Similarity for Comparison of Neural Networks"

The original idea was from a series of discussions between all authors. Tianyu Cui developed the method with Prof. Marttinen and implemented most experiments. The author proposed and implemented the transfer learning experiments. The manuscript was mainly written by Tianyu Cui with contributions from all authors.

Publication IV: "Improving Medical Multi-modal Contrastive Learning with Expert Annotations"

The problem statement was defined by Prof. Marttinen and the author. The author developed the method and conducted all the experiments with inputs from Prof. Marttinen. The author wrote the manuscript with guidance from Prof. Marttinen.

Abbreviations

BERT Bi-directional Encoder Representations from Transformers

CKA Centered Kernel Alignment

CLIP Contrastive Language Image Pretraining

EHR Electronic Health Records

GPT Generative Pretrained Transformers

NN Neural Networks

RSA Representational Similarity Analysis

1. Introduction

The central challenge in data science has shifted from the problem of managing too much data to facing a shortage of high-quality data [14, 82, 32]. This issue is particularly pronounced in sectors like healthcare where precision is crucial. Although the transition from paper-based to digital health records across the globe has significantly improved data accessibility, the quality of data remains constrained.

Electronic Health Records (EHR) capture an individual's medical history and their complex interactions with the healthcare system over time. These records are crucial for developing predictive models that can forecast health outcomes and help improve patient management strategies. This thesis leverages deep learning [55] to harness the predictive power of EHR data, addressing the challenge of data scarcity and quality in healthcare.

Neural networks, particularly large-scale models, have demonstrated exceptional performance across various domains largely due to the availability of large datasets [16, 26, 83, 54]. However, the unique constraints of healthcare data – such as its sensitivity, privacy concerns, and the need for expertly annotated data – complicates the collection of such large-scale datasets. To navigate these challenges, this thesis applies transfer learning techniques to achieve high model efficacy even with limited data availability.

The thesis explores four distinct research questions, aiming to illustrate how specialized deep learning techniques can enhance the utility of EHR data.

Forecasting Healthcare Utilization
The thesis first investigates whether the integration of temporal dynamics within EHR through the use of sequential deep learning models can improve the prediction accuracy in healthcare utilization (Publication I). We task the model with predicting the future healthcare usage of an individual using their health history. While Recurrent Neural Networks (RNNs) and other sequential models excel in handling time-series data by capturing temporal dependencies within patient visit records, traditional

non-sequential models remain dominant in the domains of healthcare utilization and risk adjustment. The study aims to determine whether leveraging the chronological sequence of medical events in EHRs can yield more precise forecast of healthcare needs, thus potentially supporting the planning and management of patient care.

Research Question 1: Can the integration of temporal dynamics of EHR data using sequential deep learning models improve prediction accuracy?

Transfer Learning with Specialized Models

Next in Publication II, we explore the potential benefit of building models specifically tailored for EHR data compared to generic natural language processing (NLP) models. The research hypothesizes that specialized models, developed to address the distinct structure and unique patterns present in healthcare data, will outperform conventional NLP models that are not optimized for such intricacies. The research assesses both predictive accuracy and sample efficiency, crucial metrics for the practical deployment of machine learning models in healthcare settings. By adapting deep learning architectures to better fit the specific needs and constraints of healthcare data, this research aims to demonstrate the superiority of customized models in clinical applications.

Research Question 2: Do models custom-built for EHR outperform generic NLP models in predictive accuracy and sample efficiency?

Covariate Adjusted Neural Network Similarity

Research Question 3 investigates whether the similarity metrics used to compare neural network representations can be made more consistent with their actual functional performance on different domains (Publication III). Traditional similarity measures like Centered Kernel Alignment (CKA [52]) and Representational Similarity Analysis (RSA [66, 53]) display superficial similarities dictated by the inherent structure of data rather than the true functional equivalences across models. The research introduces and evaluates a novel approach using covariate adjustment techniques to refine these metrics by adjusting for the input similarities. By accounting for these confounding influences of input data structure, it provides a more accurate assessment of functional similarities between neural networks, particularly in contexts such as transfer learning where understanding cross-domain capabilities is crucial.

Research Question 3: Can we make the representation similarity between

neural networks more consistent with their functional similarity by correcting for the bias caused by the input structure?

Multi-modal Contrastive Learning with Expert Annotations
Final research question in Publication IV, investigates whether integrating high-quality expert annotations into a medical CLIP model [76] can enhance its performance. The study evaluates the impact of incorporating expert-generated annotations, like radiologist eye-gaze heatmaps, on improving multi-modal learning outcomes in medical CLIP models. By more precisely aligning visual and textual data, the research aims to demonstrate how expert annotations can significantly improve the model's generalization capabilities across medical imaging tasks, thereby increasing diagnostic accuracy.

Research Question 4: Can high-quality expert annotations improve the baseline performance of a medical CLIP model?

Outline
Chapter 2 will provide the essential background needed to understand the data handling and methodologies employed in this thesis. It will cover fundamental concepts relevant to the analysis of Electronic Health Records (EHR), the techniques of transfer learning, and the methods used to compare neural network models. Chapters 3 through 6 will each summarize a publication derived from this research, detailing the methods, key results and their implications in the field of healthcare data analytics. The final chapter will offer conclusions based on findings from these individual studies and propose directions for future research, aiming to further enhance the applicability and accuracy of machine learning models in healthcare.

2. Background

This chapter covers the foundational knowledge necessary to understand the methodologies and applications discussed in this thesis. It begins with an exploration of Electronic Health Records (EHR) data, detailing its characteristics, how it is represented, and the preprocessing steps necessary for machine learning models. Following this, the chapter delves into key concepts in deep learning, particularly in the context of healthcare data, and motivates the need for developing specialized models to effectively handle this complex data.

2.1 Electronic Health Records: An Overview

The interaction of an individual with the healthcare system begins at an early age, including during the neonatal phase, and continues throughout their life. This ongoing series of interactions forms what is termed as their health trajectory. Analyzing these trajectories is crucial as it provides insights into future health conditions of a patient and patterns that could be predictive for other patients with similar health histories. The systematic collection and analysis of Electronic Health Records (EHR) are vital for such studies.

EHRs originated from the need to maintain consistent and accurate medical records. Over the years, as the methods of record-keeping evolved from paper-based to digital systems, there emerged a need for more efficient data management and accessibility across different healthcare providers [23]. EHR systems often operate within centralized databases especially within hospital groups or across healthcare facilities in many regions [88, 99]. Some countries have implemented national databases that store all citizen's health records, enhancing accessibility and continuity of care across different healthcare providers [93, 6, 49].

EHRs encompass a variety of data types essential for comprehensive healthcare management [70, 22]. They include structured data, which pertains to operational aspects such as administrative details, billing infor-

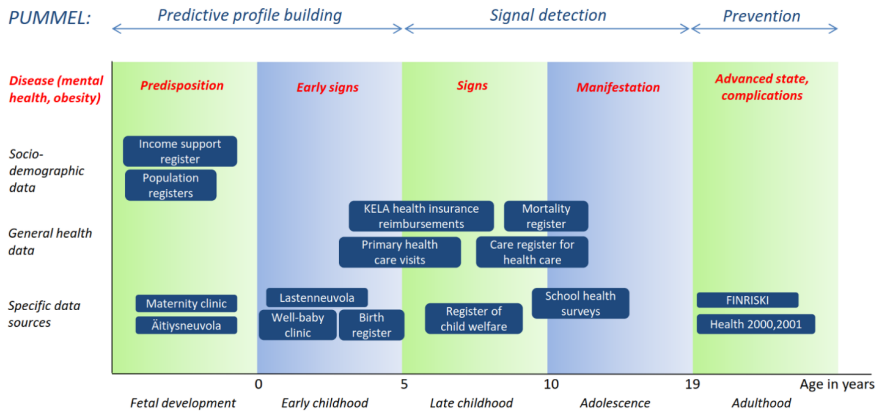

Figure 2.1. Timeline of a subset of health registries in Finland. ©Pekka Marttinen reused with permission.

mation, and patient demographics, all stored in tabular form. Clinical data, also stored in tabular format, include medical diagnoses, medications, and procedures performed during each visit. Additionally, EHRs contain time series data such as vital signs, including heart rate and blood pressure, which are crucial for ongoing patient monitoring. Moreover, EHRs incorporate unstructured data, such as free-text patient notes and radiology images, providing deeper insights into a patient's medical condition and aiding diagnostic processes.

2.1.1 Structured EHR

Structured data within EHRs primarily consist of information organized in tabular form, such as administrative, billing, and demographic details. It also includes clinical data such as diagnoses, medications, and procedures associated with each patient interaction with the health center. To standardize diagnoses and treatments further, EHRs utilize established medical coding systems such as the International Classification of Diseases (ICD) [7] and the International Classification of Primary Care (ICPC) [4]. The ICD is developed by the World Health Organization (WHO), and is a global standard. It provides a comprehensive system to classify and code all symptoms, disorders, and external causes of diseases. On the other hand, the ICPC codes patient encounters in primary care including symptoms, diagnoses, and interventions.

2.1.2 Cohort

For Publications I and II, we utilize a comprehensive dataset obtained from the Registry of Primary Healthcare Visits of the Finnish Institute for Health and Welfare (THL). Finland maintains a comprehensive col-

Table 2.1. Basic statistics of EHR dataset between the years 2012 to 2018 (Publication I & II)

# of unique patients	1,396,766	Max # of codes per visit	47
# of visits	271,103,617	# of unique ICD-10 codes	12,082
Avg. # of visits per patient	182.733	# of unique ICPC-2 codes	1,360
Avg. # of codes per visit	2.783	# of unique procedure codes	495

lection of registries pertaining to different stages of a person's lifetime (Figure 2.1). The study uses the dataset consisting of both outpatient (Avo-Hilmo) and inpatient (Hilmo) visit information for every Finnish citizen aged 65 or above, spanning 2012 to 2018. This pseudonymized dataset encompasses demographic information, diagnosis and procedure codes for each patient visit, and provides a broad view of healthcare interactions including telephone consultations, home care visits, and traditional clinic visits. The data was pseudonymized by authorized personnel at THL to ensure patient privacy. Furthermore, all data processing and computation were performed in an encrypted cloud computing infrastructure provided by CSC – IT Center for Science, Finland's "ePouta" service, adhering to strict data security protocols. Table 2.1 shows the key statistics of the dataset (Publication I & II).

2.1.3 Challenges in EHR data

The usage of EHR in health analytics comes with a few challenges, particularly in the handling of noisy, incomplete, or corrupted data. EHR data could exhibit inconsistencies due to variations in data collection practices across different healthcare providers. Additionally, the data may become corrupted during real-time collection or storage, leading to misleading patient history [29]. Addressing these issues often requires advanced data preprocessing and imputation techniques [89, 74, 41], as well as robust models capable of handling such imperfections [24, 48, 27].

2.2 Data Representation for Structured EHR

To enable machine learning analysis, the raw tabular data within EHRs must be transformed into a chronological sequence of healthcare interactions for each patient. This transformation effectively converts it into a multivariate time series. Each data point in this series is represented as a tuple (t, X_t), where t is the timestamp of the visit, and X_t includes all observations made during that visit. During the encoding process, diagnosis and procedure codes $c \in \mathcal{V}$ are transformed into one-hot-encoded vectors $\hat{c} \in \mathbb{R}^{|\mathcal{V}|}$. These vectors are then organized based on their occurance within each visit, ensuring that temporal dynamics within the patient's health

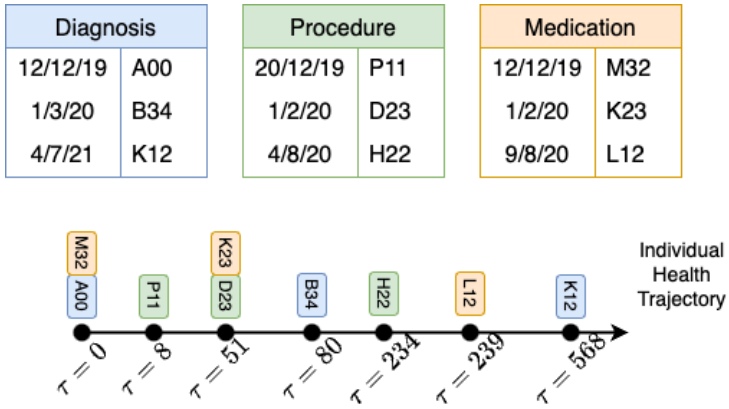

Figure 2.2. Transforming EHR into Sequential Format. This figure illustrates the conversion of tabular EHR data into structured multivariate sequential format for analysis. To manage varying visit frequencies and reduce data sparsity, we use a relative time frame, τ, where $\tau = 0$ corresponds to the patient's first interaction and subsequent visits are marked by the days elapsed since this date. ©Yogesh Kumar

trajectory are maintained and can be analyzed effectively.

In Figure 2.2, we illustrate the transformation of tabular data into a multivariate sequential format suitable for sequential analysis. This process begins by sorting the records from relevant tables based on time to ensure each patient's healthcare interactions are chronologically organized. However, the frequency of patient visits can vary significantly, with intervals ranging from several months to years. To address this potential sparsity caused by these irregular intervals, we adopt a relative time frame, τ, rather than an absolute one. Here, $\tau = 0$ is assigned to the first recorded interaction of the patient. Subsequent visits are then timestamped with the number of days elapsed since this initial interaction. This relative time frame approach helps to reduce the sparsity while maintaining the temporal information contained in the health trajectory.

2.2.1 Representation for Non-sequential Models

For analysis with non-sequential models, Publication I employs two count-based approaches as baseline: Lasso Regression and Gradient Boosted Decision Trees (GBDT). To prepare the structured EHR data for these non-sequential models, we utilize TF-IDF (Term Frequency-Inverse Document Frequency) encoding. This technique converts categorical data into a numerical format that is more suitable for model input. Unlike simpler bag-of-words representations, TF-IDF improves model performance by reducing the weight of terms that appear frequently across different patient sequences. This is achieved by adjusting the term frequency for each token

based on its document frequency. It follows the formula:

$$\text{TF-IDF}(t,D) \in \log \frac{N}{|d \in D : t \in d|} \tag{2.1}$$

where D is the dataset of patient sequences, N is the total number of sequences, and $|d \in D : t \in d|$ counts sequences containing the token t. This adjustment allows our non-sequential models to discern patterns without being biased by the prevalence of certain terms.

2.2.2 Representation for Sequential Models

To analyze sequential medical data effectively, we adapt techniques from Natural Language Processing (NLP). This involves employing an embedding layer that transforms one-hot-encoded (OHE) vectors into a denser D-dimensional vector space. Specifically, each OHE vector, representing discrete medical tokens such as diagnosis and procedure codes, is multiplied by an embedding matrix $W_{emb} \in \mathbb{R}^{|\mathcal{V}| \times D}$, where $\mathcal{V}$ is the vocabulary set of all unique codes present in the dataset. After this projection, the input is reshaped to $T \times W \times D$, where T represents the total number of visits and W indicates the number of different medical codes recorded per visit.

Sequential neural network models, such as Recurrent Neural Networks (RNNs), typically require input dimensions of $T \times D$. Hence, the data currently structured as $T \times W \times D$ needs to be aggregated over the intra-visit axis (W). Common methods for this aggregation include averaging or summing along the W axis [12]. However, more sophisticated methods could also be employed to capture intricate relationships within visits, potentially improving model performance [11, 58].

2.3 Sequential Models for Structured EHRs

Having processed the structured EHR into a format compatible with sequential analysis, we turn our focus to various sequential neural network models that are commonly employed to handle this type of data. Historically, the domain has predominantly utilized Recurrent Neural Network (RNN) based models, such as Long Short-Term Memory (LSTM) networks [60, 11] and Gated Recurrent Units (GRU) [12], which are well-suited for handling data with temporal dependencies. More recently, there has been a shift towards incorporating transformer-based models [92], which offer advantages in processing complex and long-range dependencies [79, 68, 56, 50].

2.3.1 Long Short-Term Memory (LSTM) Networks

The LSTM (Long Short-Term Memory) network [35] is renowned for its efficacy in managing long-term dependencies. LSTM incorporates a series of gates – input (i_t), forget (f_t), and output (o_t) – alongside cell state updates (C_t) to modulate the flow of information, as detailed in the following equations:

$$i_t = \sigma(X_t U^i + h_{t-1} W^i) \tag{2.2}$$

$$f_t = \sigma(X_t U^f + h_{t-1} W^f) \tag{2.3}$$

$$\tilde{C}_t = \tanh(X_t U^g + h_{t-1} W^g) \tag{2.4}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.5}$$

$$o_t = \sigma(X_t U^o + h_{t-1} W^o) \tag{2.6}$$

$$h_t = \tanh(C_t) * o_t \tag{2.7}$$

In the context of EHR data, the hidden states (h_t) can be interpreted as capturing the evolving health trajectory of a patient over time. This model is particularly adept at processing EHR data due to its ability to handle sequences with variable lengths and time gaps. Depending on the application, the LSTM can be configured to produce an output at each time-step or only a single output at the end of the sequence. For tasks requiring an understanding of the entire input sequence, such as patient risk stratification, or treatment outcome prediction, employing a bidirectional LSTM [84] might be advantageous since it allows the model to gather context from both past and future states.

2.3.2 Multi-headed Attention

Multi-headed attention (MHA) is a mechanism used in neural networks to enhance the model's ability to capture complex relationships within data. Originally designed for natural language processing, it has been effectively adapted for various other applications with sequential data including the analysis of EHRs. The concept involves using self-attention mechanism, which enables the model to assign attention scores to each input token to modify their importance while making prediction. MHA can be described mathematically as:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \times V \tag{2.8}$$

Here, Q,K and V represent the query, key and value matrices, respectively, and d represents the dimension of the vectors. The input, X, is transformed into query, key and value by multiplying it with trainable matrices W_q, W_k

and W_v, respectively. The normalization by $\sqrt{d}$ is crucial as it prevents the softmax function from saturating when the inner products are large in magnitude.

The Transformer model [92] consists multiple network layers, each containing the MHA mechanism, a feed forward mechanism and Layer Normalization [1] to create a robust framework for processing sequential data. For a comprehensive overview of the Transformer architecture and its variants refer to (Rush, 2018) [81] and (Mary and Hutter, 2022) [73].

2.4 Pre-training Neural Networks

Pre-training is an important technique in deep learning aimed at improving the performance, especially when dealing with large models and relatively small datasets. This method involves initially training a network on a different, typically larger dataset (D_S) before fine-tuning it on the target dataset (D) and for task (T). During the fine-tuning phase, the model benefits from being initialized with weights that are not random, which leads to better convergence and effective learning on smaller datasets [21]. Further, models that undergo pre-training have demonstrated improved robustness and more reliable uncertainty estimates in predictions [33]. This process is closely related to transfer learning, where the skills learned during pre-training on one dataset or task are transferred to improve performance on another dataset or task [28].

Transfer Learning from ImageNet
The strategy of training a model on a large dataset (D_S) for a similar task (T_S) is widely utilized in computer vision. A notable example is the use of ImageNet, a large dataset consisting of over 14 million images manually categorized into approximately 20,000 labels [16]. Training on ImageNet enables the models to learn general features beneficial across several visual recognition tasks. The extent to which the layers of the pre-trained network are frozen during the fine-tuning process depends on the similarity of between the target task (T) and source task (T_S).

Transfer Learning in Healthcare
In healthcare, models pre-trained on ImageNet have been successfully fine-tuned for medical imaging tasks, such as detecting diseases in chest X-rays [47]. Beyond medical imaging, transfer learning has also been applied in natural language processing (NLP) for electronic health records (EHRs), enabling tasks such as clinical named entity recognition and medical text classification [58, 79]. These approaches have demonstrated improved performance, particularly in data-scarce situations, by leveraging the general visual or textual features learned during pre-training on large,

diverse datasets.

Self-supervised learning
Self-supervised learning creates its own labels by using inherent structures or predictions within the data itself. This is different from unsupervised learning which involves learning patterns without any explicit labels. By leveraging the input data to generate labels, it allows learning useful representations without manually annotated labels [2].

In natural language processing, self-supervised learning forms the backbone of pre-training strategies in models such as BERT (Bidirectional Encoder Representations from Transformers) [17, 62] and GPT (Generative Pre-trained Transformer) [77, 9]. Similarly, in computer vision, methods like SimCLR (Simple Framework for Contrastive Learning of Visual Representations) [10] and Barlow Twins [103] employ contrastive learning techniques for self-supervision. Building on these concepts, our research in Publication IV expands into multi-modal domain with the contrastive pre-training of image and text data using CLIP (Contrastive Language-Image Pre-training). We will explore CLIP in greater detail in the following section.

2.4.1 Contrastive Language-Image Pre-training (CLIP)

CLIP extends the principles of contrastive learning to the multi-modal domain, specifically between images and text [76]. The model consists of two primary components: an image encoder and a text encoder. It is trained on a vast dataset of image-text pairs sourced from the internet, such as images alongside their corresponding captions [100]. The image encoder transforms visual inputs into feature vectors (or embeddings), while the text encoder converts the textual descriptions into text feature vectors. The training objective is to maximize the cosine similarity between vectors of matching image-text pairs and minimizing the similarity between the non-matching pairs. In healthcare setting, CLIP's application involves analyzing medical scans alongside descriptive reports, such as pairing chest X-rays with radiology reports prepared by healthcare professionals [105, 37, 97, 101].

Concretely, we can consider a chest X-ray image I_i and its radiology report T_i, indexed by i in the dataset. We can represent the image and text embeddings which reside in a d-dimensional space, as v_i and t_i, respectively, with $v_i, t_i \in \mathbb{R}^d$. The image embeddings v_i are obtained via an encoder $f(I_i)$, and the text embeddings t_i through $g(T_i)$. CLIP uses the InfoNCE loss [72], which is defined as follows:

$$\mathcal{L}_{\text{image}} = \mathop{\mathbb{E}}_{(v_i,t_i)\sim pos}\left[-\log \frac{\exp(sim(v_i,t_i)/\tau)}{\exp(sim(v_i,t_i)/\tau) + \sum_{j\neq i}\exp(sim(v_i,t_j)/\tau)}\right] \tag{2.9}$$

$$\mathscr{L}_{\text{text}} = \mathop{\mathbb{E}}_{(t_i,v_i)\sim pos}\left[-\log \frac{\exp(sim(t_i,v_i)/\tau)}{\exp(sim(t_i,v_i)/\tau)+\sum_{j\neq i}\exp(sim(t_i,v_j)/\tau)}\right] \tag{2.10}$$

The total loss is then the average of the losses computed for both image-to-text ($\mathscr{L}_{\text{image}}$) and text-to-image ($\mathscr{L}_{\text{text}}$) mappings, expressed as $\mathscr{L}_{\text{total}} = \frac{1}{2}\left(\mathscr{L}_{\text{text}} + \mathscr{L}_{\text{image}}\right)$. Here, τ is the temperature parameter to control the scale of the similarity scores. Its value is typically set to be a learnable parameter during training. The loss expectation is taken over all the positive pairings in the dataset.

Theoretically, two key concepts underpin the quality of embeddings produced by contrastive learning models: alignment and uniformity [95, 94]. **Alignment** refers to the model's ability to reduce the distance between embeddings of positive pairs – such as an image and its corresponding caption – thereby ensuring that semantically similar data points are close in the embedding space. This is crucial for the model's effectiveness in correctly associating related image-text pairs. On the other hand, **uniformity** addresses the distribution of embeddings across the unit hypersphere. By encouraging a uniform distribution, the model prevents the embeddings from clustering too tightly, which could limit its ability to generalize to new, unseen data. Formally, these concepts are captured as follows [71]:

$$\text{Alignment} = -\mathbb{E}_{(v_i,t_i)\sim pos}\left[\|v_i - t_i\|_2^2 - \min_{j\neq i}\|v_i - t_j\|_2^2\right] \tag{2.11}$$

$$\text{Uniformity} = -\log \mathbb{E}_{(v_i,t_j)\sim \mathscr{D}}\left[\exp(-2\|v_i - t_j\|_2^2)\right] \tag{2.12}$$

The contrastive pre-training task not only improves the model's accuracy in identifying relevant associations between images and text but also broadens its capabilities across several downstream tasks, such as zero-shot classification and retrieval for both images and text. This versatility of CLIP has opened up new possibilities for these embeddings in various AI applications including in healthcare domain [106, 19, 104]. In Publication IV, we explore potential enhancements to CLIP by integrating additional high-quality labels.

2.5 Measuring Similarities between Neural Networks

Centered Kernel Alignment (CKA [52]) and Representational Similarity Analysis (RSA [53]) are two key methodologies for evaluating the similarity of neural network representations. CKA quantifies similarity between models by measuring the alignment of activation patterns in response to the same inputs, indicating which layers perform comparable functions or capture similar features. RSA on the other hand, achieves this by comparing the dissimilarity matrices, which are constructed from the patterns of neural activations in response to various inputs, to reveal

how different networks process these inputs and highlight functional differences between them [52, 66, 53, 85, 69, 51, 78].

To illustrate, consider $X \in \mathbb{R}^{n \times p}$ as the input dataset with n data points and p features. Representations from two neural networks at given layers are denoted by $X_{f_1}^{m_1} \in \mathbb{R}^{n \times p_1}$ and $X_{f_2}^{m_2} \in \mathbb{R}^{n \times p_2}$, for the first and second network respectively, where p_1 and p_2 represent the dimensions of the respective layers. The matrices are centered by subtracting the feature mean for each element, and normalized by dividing each element by the Frobenius norm of the matrix to facilitate comparison.

A common method for comparing neural network representations involves assessing the similarity structures of each network. This process begins by calculating the pairwise similarity among examples in the representations from two neural network layers, $X_{f_1}^{m_1}$ and $X_{f_2}^{m_2}$, using a similarity function $k(\cdot,\cdot)$:

$$K_{f_1}^{m_1} = k(X_{f_1}^{m_1}, X^{m_1} f_1), \quad K_{f_2}^{m_2} = k(X_{f_2}^{m_2}, X_{f_2}^{m_2}), \tag{2.13}$$

where $K_{f_1}^{m_1}$ and $K_{f_2}^{m_2}$ are called the representational similarity matrices (RSMs). These matrices capture the internal similarity within each layer's outputs. To quantify the similarity between these two neural network layers, a secondary similarity measure, $s(\cdot,\cdot)$, is applied:

$$s_{f_1,f_2}^{m_1,m_2} = s(K_{f_1}^{m_1}, K_{f_2}^{m_2}), \tag{2.14}$$

which evaluates how similarly the networks process their inputs.

CKA uses a kernel function to establish initial similarities, $k(\cdot,\cdot)$, and applies the Hilbert-Schmidt Independence Criterion (HSIC) for the second, $s(\cdot,\cdot)$. RSA, in contrast, measures inter-example dissimilarities using Euclidean distance and evaluates these distances using Pearson or Spearman correlations [66, 85]. In Chapter 5 we improve upon these measures by performing covariate adjustment for the input.

3. Forecasting Healthcare Utilization

One critical application of Electronic Health Records (EHR) analysis is in forecasting healthcare utilization, which is pivotal for resource allocation and planning within healthcare systems [20]. Accurate predictions of future healthcare costs at an individual level enable a more efficient and equitable distribution of medical resources. Traditional models often rely on basic regression techniques that utilize demographic and previous year's diagnosis counts [80, 87, 20, 8]. This chapter summarizes Publication I, where we explore whether the inclusion of temporal information, using sequential models, would improve the efficacy of these models.

Research Question 1: Can the integration of temporal dynamics of EHR data using sequential deep learning models improve prediction accuracy?

This question addresses the potential of sequential models, such as Recurrent Neural Networks (RNNs) with multi-headed attention, to enhance predictive accuracy by capturing the temporal patterns inherent in patients' healthcare trajectories. We develop a model that uses the number of physical visits to a General Practitioner (GP) as a proxy for upcoming healthcare costs.

3.1 Method

We employ a sequential deep neural network to forecast healthcare utilization using patient history. At the core of our model is a stack of RNN layers, specifically Long Short-Term Memory (LSTM) units, known for their effectiveness in handling long-term dependencies in the input sequence. The raw EHR input is transformed using the steps described in Section 2.2.2. Each patient visit i.e., encapsulated as a tuple (t, X_t), where t is the timestamp and X_t contains the observations (diagnosis and procedure codes) for the t^{th} visit.

In the transformed EHR data, each patient visit, represented as a time-

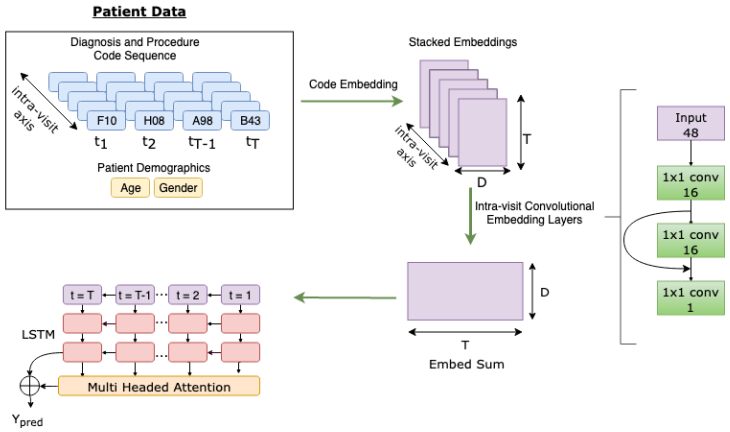

Figure 3.1. This diagram illustrates a deep sequential model applied to longitudinal individual health records. It includes LSTM layers for handling time-sequenced data, a convolutional embedding block to combine intra-visit codes, and a Multi-Headed Attention (MHA) layer designed to improve the model's ability to focus on significant long-term interactions in the patient data. (Publication I)

step t, may involve multiple clinical codes. These can vary greatly in number, with as many as 47 different codes per visit observed in our dataset (Table 2.1). To efficiently proess these multiple codes, the embeddings generated from the embedding layer for each code need to be aggregated along the intra-visit axis, j. While previous studies have often simplified this aggregation by summing the embeddings across j [12], our approach considers a more sophisticated approach. As shown in Figure 3.1, the convolutional embeddings block employs a sequence of 1×1 convolution operations specifically designed to modify the size of the input tensor along the intra-visit axis, while preserving dimensions along other axes. This block includes three 1×1 convolutions, interspersed with batch normalization [38], PReLU activation [30] and a skip-connection [31] between the first and third layers. The sequence relative time variable, τ (see Section 2.2), for each time step is then concatenated to the embedding matrix.

The model also integrated a multi-headed attention (MHA) layer for capturing the long-term interactions in the patient health trajectories. After processing the sequence through LSTM layers, the hidden stares are subjected to MHA, where the keys, values and queries are all derived from the LSTM hidden states. The attention outputs are then aggregated using either average or max pooling. The resulting vector is then passed through two fully-connected layers to the output layer, which finalizes the prediction. This combination allows the model to enhance its sensitivity to significant events in the patient's history, even if they occurred in the distant past. Figure 3.1 shows the schematic of the entire sequential model architecture. The source code is available online.

Table 3.1. Test R^2 score for year 2016. The LSTM model outperforms the non-sequential models used as baseline. We can also see that the performance improves as we increase the context length of the input. (Publication I)

Training Year(s)	Lasso	LightGBM [46]	LSTM_{mh_attn}
2015	0.0367	0.2854	**0.2974**
2014, 2015	0.0752	0.3094	**0.3259**
2013, 2014, 2015	0.0913	0.3248	**0.3436**

3.2 Key Results

In this study, we aimed to predict the number of physical visits to healthcare centers in the following year using patient EHR data from previous years, specifically analyzing the clinical code sequences $x_1, x_2, \ldots, x_T$ and patient features x_{num}. We focus solely on physical visits, which include outpatient services, hospital visits, and receptions at health centers. This approach allows us to target resource allocation where it is most needed, since these visits demand more resources compared to other less resource-intensive interactions, such as phone consultations.

Prediction accuracy improves with longer context
In our analysis, we assessed the performance of a sequential deep learning model against non-sequential baselines (Lasso and LightGBM [46]). We focused on the coefficient of determination, R^2 scores, to evaluate predictive accuracy for healthcare utilization. To verify the robustness of the model's performance, the experiment was repeated with an additional run using a different random seed. The results in Table 3.1 demonstrate that the sequential model consistently outperforms the baseline models across all training periods. Additionally, the improvement in R^2 scores with increasing data context length indicates that a more extensive historical context significantly enhances the model's forecasting capability. This highlights the advantages of sequential models in capturing temporal dynamics in patient data.

Sequential model generalizes to future years
In practical applications, there is typically a time lag between the training phase and the operational use of risk adjustment model. Hence the machine learning models are expected to generalize to future years without the need for retraining [20]. To evaluate our model's robustness in this aspect, we trained using data from 2014 to 2015 to predict the healthcare visits in 2016. We then test the model's prediction for 2017 using data from 2015-2016 and for 2018 using data from 2016-2017. As illustrated in Figure 3.2, the sequential models, due to their ability to capture tempo-

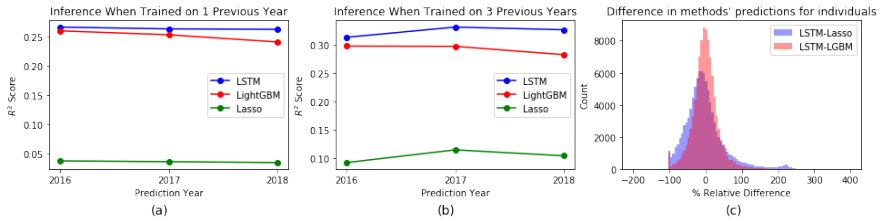

Figure 3.2. *Left, Center*: R^2 score when inferred on future years (all models were trained with $100K$ examples to predict 2016, no data from 2017 or 2018 was used in training); *Right*: Histogram showing the % relative change in predicted visits if we switch from Lasso or LightGBM to an LSTM model (Publication I)

ral dependencies, outperform the count-based models in generalization to future years.

Finally, we assess the practical impact of using our LSTM model for healthcare resource allocation (Figure 3.2 *Right*), which is based on predicted costs for registered individuals. Although overall improvement in R^2 score is about 10%, individual predictions vary significantly, suggesting substantial effects on resource distribution.

Limitations and Challenges

Our method incurs higher computational costs due to the complexity of deep neural networks compared to traditional models like Lasso and LightGBM. Additionally, neural networks such as RNNs inherently lack interpretability. Therefore, in scenarios where interpretability is crucial, traditional models may be more suitable [36].

3.3 Revisiting Research Question 1

In addressing *Research Question 1*, our goal was to determine whether incorporating the temporal dynamics of EHR data through sequential deep learning models could enhance prediction accuracy. We found that our LSTM model, augmented with multi-headed attention, consistently outperformed traditional non-sequential count-based models such as Lasso and LightGBM in forecasting healthcare utilization. The results not only showed improved R^2 scores but also demonstrated the model's ability to effectively handle data spanning multiple years, indicating strong generalization capabilities. Based on these findings, we conclude that integrating sequential deep learning models into EHR data analysis is advantageous and warrants further exploration to improve predictive accuracy in healthcare applications.

4. Improving Sample Efficiency with Transfer Learning

Deep learning has had a significant impact in the natural language processing (NLP) domain, and due to the structural similarities between NLP and sequential EHR data, these models have consequently been applied to EHR analysis as well [12, 58, 60, 68]. However, applying deep learning models to EHR data presents unique challenges.

Unlike the coherent sequences typically found in NLP tasks, healthcare data often consists of sparse and irregularly timed clinical events, which adds complexity to its analysis. Moreover, the effective training of these models frequently requires vast amounts of labeled data, a resource not always available in healthcare domains. To address this, researchers often resort to pre-training strategies like Masked Language Modeling (MLM) or autoregressive techniques to enhance model performance on downstream tasks [17, 77].

Another significant challenge arises with the need to cater to divergent patient subgroups. Patients diagnosed with specific diseases often exhibit health trajectories that are distinct from that of the general population (Figure 4.1). This necessitates the use of customized predictive models for each group. However, the data scarcity in these subgroups poses problems, making the training of robust, large-scale models impractical [87, 20].

This chapter summarizes Publication II, which seeks to explore how EHR data, consisting of sequence of clinical codes, diverges from the data in traditional NLP applications, and how these differences require the development of tailored models and pre-training methods. Our research question focuses on whether specialized methods can overcome the limitations of conventional "off-the-shelf" approaches borrowed from other domains like NLP.

Research Question 2: Do models custom-built for EHR outperform generic NLP models in predictive accuracy and sample efficiency?

The subsequent sections will introduce our model, SANSformers, designed specifically for the challenges of EHR, and discuss the implementa-

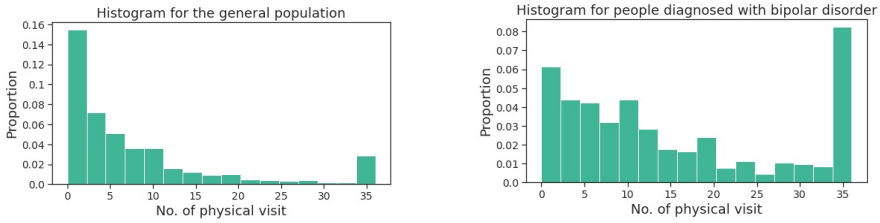

Figure 4.1. Illustrating the Challenge of Predicting for Diverse Subgroups. The histograms show hospital visit frequencies for two distinct groups: the general population (*Left*) and individuals with bipolar disorder (*Right*). These visualizations reveal significant differences in visit patterns among subgroups, underscoring the challenges in forecasting healthcare utilization (Publication II)

tion of Generative Summary Pre-training (GSP) to effectively harness the potential of self-supervised learning in this domain.

4.1 Method

In recent years, the Transformer architecture with minor modifications has been adapted for EHR analysis [68, 79, 58, 50]. However, their core self-attention mechanism may not be optimal due to the sparse and irregular nature of EHR. Recent developments suggest that simpler, more computationally efficient structures, such as all-MLP (Multi Layer Perceptron) models, could replace self-attention in various NLP and computer vision applications [90, 61, 91, 57, 67]. Inspired by this, we develop SANSformers, a model specifically designed for EHR data that forgoes traditional self-attention to better suit the requirements of healthcare data analysis. The SANSformer retains key features from traditional Transformer models—positional encodings, skip-connections [31], and layer normalization [1]. The architecture of the SANSformer model comparing the Mixer mechanism with self-attention is shown in Figure 4.2.

Adapting Axial Decomposition for EHR

As outlined in Section 2.2.2, the processed EHR data is of dimension $T \times W \times D$ where T represents the total number of visits and W indicates the intra-visit dimension. Traditional methods that flatten the input sequence lead to computational complexities scaling as $\mathcal{O}(T^2W^2)$. To mitigate this, we implement axial decomposition [34] which involves decomposing the mixer or attention mechanism into row-wise and column-wise operations. Specifically, it transforms the original tensor of dimension $T \times W \times D$ into T tensors of $W \times D$ and W tensors of $T \times D$, thereby reducing complexity to $\mathcal{O}(TW^2 + T^2W)$.

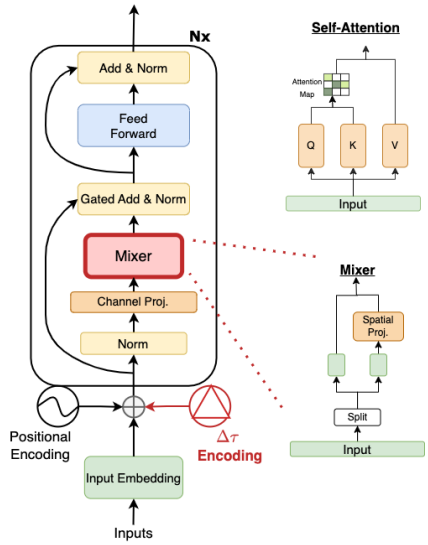

Figure 4.2. SANSformer Architecture. This figure illustrates a single layer of the SANSformer model. We present modifications to the conventional Transformer architecture by Vaswani et al. [92], highlighting changes such as the inclusion of $\Delta\tau$ embeddings and the attention-free mixers which replace self-attention. A comparison on the right showcases the difference between conventional self-attention and Mixer, demonstrating the efficiency achieved with fewer projection operations (Publication II)

Attention-less Mixers for EHRs

After the axial decomposition of EHR data, each patient's data is represented by two matrices: $X_t \in \mathbb{R}^{T\times D}$ for temporal tokens across visits, and $X_v \in \mathbb{R}^{T\times D}$ for tokens within the same visit. The SANSformers model transforms these matrices using non-linear operations referred to as "mixers", which replace the traditional self-attention mechanism used in Transformer models. Specifically, these mixers perform transformations similar to 1×1 convolutions, involving matrix multiplications that facilitate cross-token interactions. This mixing process is defined by the following operations (described for X_t):

$$Z = \text{GELU}(X_t U) \tag{4.1}$$

$$\widehat{Z} = \text{SGU}(Z) \tag{4.2}$$

$$Y = \widehat{Z} V \tag{4.3}$$

where $X_t \in \mathbb{R}^{T\times D}$ represents the input matrix with T denoting sequence length and D the embedding dimension. $U \in \mathbb{R}^{D\times 2P}$ and $V \in \mathbb{R}^{P\times D}$ are trainable weight matrices, with P as the projection dimension.

The Spatial Gating Unit (SGU) further processed Z, which is split into two halves Z_1 and Z_2, and an affine transformation is applied to Z_2:

$$Z_1, Z_2 = \text{split}(Z) \tag{4.4}$$

$$\widehat{Z_2}^T = \text{GELU}(Z_2^T W + b) \tag{4.5}$$

$$\widehat{Z} = Z_1 \odot \widehat{Z_2} \tag{4.6}$$

Here, $W \in \mathbb{R}^{T \times T}$ adjusts for causal dependencies by zeroing out all upper-triangular elements, ensuring the model remains autoregressive. This structure allows for efficient cross-time interaction of the tokens without relying on the computationally intense self-attention mechanism, demonstrating that mixers provide a viable, simpler alternative for managing the complexities of EHR data.

Encoding Temporal Information for EHR
Similar to Transformers, SANSformers incorporate positional encodings using sinusoidal functions to uniquely identify the position of the tokens in the input sequence. However, two visits which have a large time gap between them are typically unrelated. To address this, the model additionally includes $\Delta\tau$ embeddings, which capture the number of days between each consecutive visit.

Generative Summary Pre-training (GSP)
The Generative Summary Pre-training (GSP) is a transfer learning strategy designed to enhance predictive accuracy for specific patient subgroups by utilizing historical data from the general patient population. Unlike in common NLP pre-training techniques such as Masked Language Modeling (MLM) or Generative Pre-trained Transformer (GPT), which focus on making token level prediction, GSP targets the forecasting of healthcare utilization.

This approach involves pre-training the model on a broad dataset that excludes patients from the subgroups from the fine-tuning phase, effectively preventing data leakage and ensuring model integrity. By focusing solely on general population data, GSP predicts the number of patient visits in the upcoming year by analyzing data from the previous year. This method not only allows GSP to leverage a diverse range of data but also enhances the model's ability to generalize better during the fine-tuning phase on targeted patient subgroups.

The source code is available online.

4.2 Key Results

The experiments utilized two EHR data sources – the Pummel dataset, sourced from registries maintained by Finnish Institute for Health and

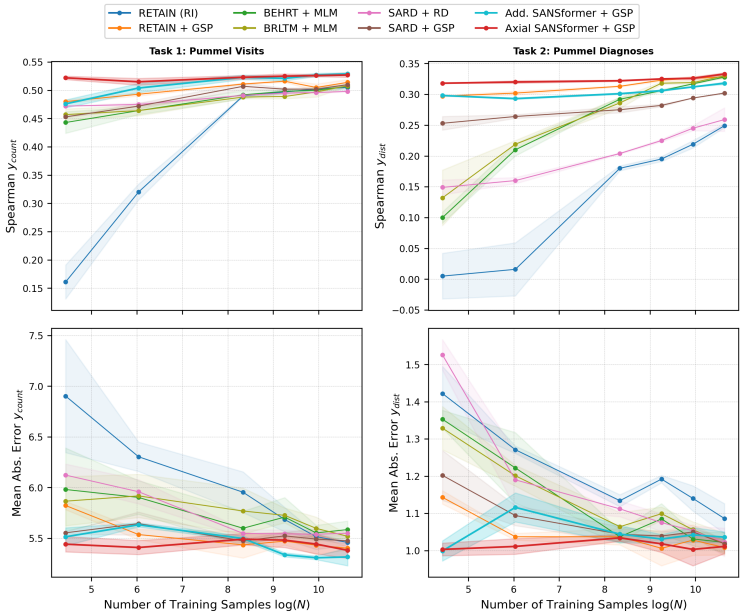

Figure 4.3. Data Efficiency through Pretraining on the Pummel Dataset. This figure assesses the effectiveness of pretrained models across various sizes of training data from the type 2 diabetes subgroup. *Row 1* displays Spearman's rank correlation for the Pummel Visits and Pummel Diagnoses tasks, while *Row 2* shows the Mean Absolute Error (MAE) for these tasks. Each data point includes the mean and standard deviation from three random restarts. Performance comparisons include the randomly initialized RETAIN (RI) model for reference (Publication II)

Welfare (THL), and the publicly available MIMIC-IV dataset [43]. For the Pummel dataset the we focused on predicting two key variables for the upcoming year based on one-year patient history:

- **Task 1 - Pummel Visits:** Predicting the number of physical healthcare center visits (y_{count}).

- **Task 2: Pummel Diagnoses:** Estimating the counts of visits for six major disease categories, significant consumers of healthcare resources in Finland. These categories include diseases such as cancer, metabolic disorders, and cardiovascular illnesses (y_{dist}).

  For the MIMIC dataset:

- **Task 3 - MIMIC Mortality:** Predicting inpatient mortality (y_{death}), using the `hospital_expire_flag` from the MIMIC-IV `admissions` table as an indicator (y_{death}).

  We evaluate our model by benchmarking it against established models

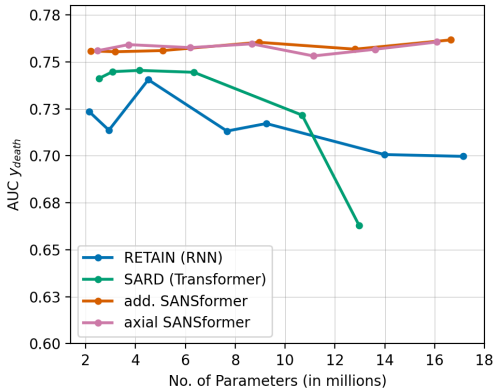

Figure 4.4. Parameter Scalability on MIMIC. This figure demonstrates the relationship between model performance on Task 3 and the number of trainable parameters, highlighting the scalability of each model. Notably, the SANSformer models exhibit consistent performance even with increased parameter counts (Publication II)

such as RETAIN [11], BEHRT [58], SARD [50], and BRLTM [68] across both the Pummel and MIMIC datasets. We also highlighted the effectiveness of our Generative Summary Pre-training (GSP) approach, particularly in enhancing prediction accuracy for specific patient subgroups within the Pummel dataset, including those diagnosed with Type 2 Diabetes, Bipolar Disorder, and Multiple Sclerosis.

SANSformers with GSP improves sample efficiency
To assess sample efficiency, we varied the training dataset size from the Type 2 Diabetes (T2D) subgroup, focusing on how the model performs with increasingly smaller datasets – an important measure given that smaller datasets typically challenge neural network performance [28]. The results depicted in Figure 4.3 illustrate that the pre-trained axial SANSformer model exhibits superior data efficiency. The figure also shows the mean and standard deviation confidence intervals from runs with three different random seeds, highlighting the model's consistency across different initializations. This model outperforms others in both the Pummel Visits and Pummel Diagnoses tasks, confirming its effectiveness even with limited data. This capability is crucial for applications in healthcare settings where expansive data may not always be available, for example in subgroups for rare diseases.

SANSformers shows better scaling trends
To examine how our SANSformers scale with an increasing number of parameters, we conducted experiments using the MIMIC dataset and Task 3, focusing on systematically increasing the size of the model by adding more layers. We selected the best-performing models from each

base architecture, comparing SANSformers against established models like RETAIN (RNN-based) and SARD (Transformer-based).

The results, illustrated in Fig. 4.4, show that SANSformers scale better in terms of parameter efficiency than their traditional Transformer and RNN counterparts. This suggests that SANSformers not only manage increased complexity effectively but also exhibit enhanced generalization capabilities, indicating robustness across varying model sizes and configurations.

Limitations and Challenges

The SANSformer models possesses certain limitations. Firstly, SANSformers are constrained by sequence length due to the dimensions of the weight matrix, which limits their effectiveness in processing very long sequences. Additionally, while the model leverages inpatient encounters to predict healthcare utilization, it relies on proxies and does not directly compute healthcare costs, which may require further adaptations for cost estimation applications [80]. Finally, the exclusion of the self-attention mechanism reduces the model's interpretability, a feature often provided by attention maps [40, 98]. Future work could explore integrating model-agnostic interpretability techniques, such as SHAP [63], to better understand the decision-making process of SANSformers in healthcare contexts.

4.3 Revisiting Research Question 2

In tackling *Research Question 2*, we explored whether a model tailored specifically for EHR domain could outperform generic techniques adapted from other fields. Our findings show that the SANSformer models, designed with unique adaptations for the complexities of EHR data demonstrate superior performance with respect to data efficiency and parameter scalability. These results support the hypothesis that domain-specific model adaptations are crucial for managing the intricacies of EHR data, leading to more effective and reliable analytical outcomes compared to generalized models from other domains.

5. Covariate Adjusted Representation Similarity for Transfer Learning

Despite the robust performance of neural networks across various domains, they are often called "black boxes" since their internal workings and decision-making processes are hard to understand. Similarity metrics such as Centered Kernel Alignment (CKA [52]) and Representational Similarity Analysis (RSA [53]) have been developed to address this issue. These metrics help compare activations between network layers, revealing which layers have the greatest influence on the model behavior (See Section 2.5 for details).

However, these conventional similarity metrics might provide misleading results when applied to models trained on different data domains. For example, CKA could indicate high similarity between two models even if their predictive performance differs significantly. This can make it challenging to understand how models share or differ in representations across domains, which is crucial in transfer learning scenarios with limited data availability. Publication III aims to modify the similarity metrics to counter this discrepancy.

Research Question 3: Can we make the representation similarity between neural networks more consistent with their functional similarity by correcting for the bias caused by the input structure?

Figure 5.1 *(Left)* demonstrates this discrepancy when applying the CKA similarity measure to the first layer activations of a ResNet-18 model using 20 random samples from the CIFAR-10 test set. This involves comparing pairs of ResNets, where one has undergone random noise addition to the pretrained weights on ImageNet. We identify pairs of examples with both high and low similarity in their input space. Remarkably, the CKA values from this perturbed ResNet still mirror the input domain similarities. This supports our hypothesis that traditional similarity metrics are heavily influenced by the input data structure. To address this, we suggest a covariate adjustment method to refine these metrics and eliminate misleading

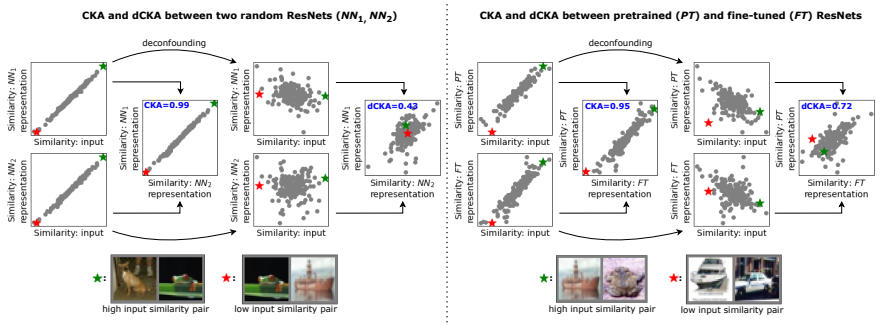

Figure 5.1. Demonstration of the Confounder in CKA. CKA measures the similarity between representations based on the inter-example similarities, which are influenced by similarities in the input space. This results in pairs with high (★) and low (★) input similarities also showing similar levels of representation similarities in both randomly initialized neural networks (*Left*) and trained neural networks (*Right*). This confounding effect leads to the counterintuitive finding that CKA values are higher in random neural networks than in pretrained and finetuned networks on similar domains (Publication III)

effects caused by spurious resemblances in the data.

5.1 Method

We propose a method, called dCKA, to improve similarity metrics by removing confounders linked to input data structures. This method employs covariate adjustment regression to eliminate the impact of input similarity from the representation similarities. By "regressing out" these confounders, we ensure that our similarity measurements truly reflect the characteristics of the model, not artifacts caused by the structure of the input data.

Using the same notations from Section 2.5, the core of our method involves adjusting the representation similarity matrices $K_{f_1}^{m_1}$ and $K_{f_2}^{m_2}$ for m_1th and m_2th layers of the networks f_1 and f_2, respectively. This adjustment removes the influence of the input similarity matrix K^0. The adjusted or deconfounded similarity matrices are given by:

$$dK_{f_1}^{m_1} = K_{f_1}^{m_1} - \hat{\alpha}_{f_1}^{m_1} K^0,\ dK_{f_2}^{m_2} = K_{f_2}^{m_2} - \hat{\alpha}_{f_2}^{m_2} K^0, \tag{5.1}$$

where $\hat{\alpha}_{f_1}^{m_1}$ and $\hat{\alpha}_{f_2}^{m_2}$ are the regression coefficients that minimize the Frobenius norm of $dK_{f_1}^{m_1}$ and $dK_{f_2}^{m_2}$ respectively. These coefficients arise from a linear regression model, which assumes that the influence of the input similarity on the representation similarity as linear and additive:

$$\mathrm{vec}(K_f^m) = \alpha_f^m \mathrm{vec}(K^0) + \boldsymbol{\epsilon}_f^m, \tag{5.2}$$

where $\mathrm{vec}(\cdot)$ flattens a matrix to a vector. Noise $\boldsymbol{\epsilon}_f^m$ is assumed to be independent of the confounder. The regression coefficients are determined

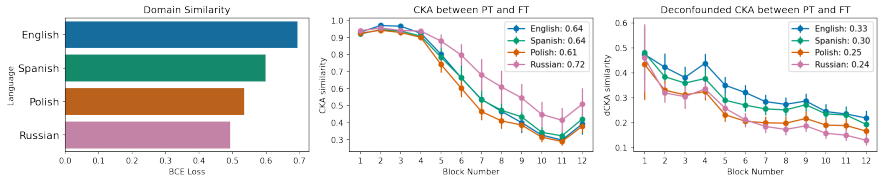

Figure 5.2. dCKA Adjusts for Transfer Learning under Domain Shift. *Left* side displays the actual domain similarity (between English and other languages) as measured by the binary cross entropy (BCE) loss in a cross-domain classifier. We illustrate the differences in similarity metrics with CKA (*Center*) and dCKA (*Right*) comparisons between the pretrained XLM-RoBERTa model and models finetuned on the STS-B task for various languages. The average similarities for each language across layers are listed in the legend. Notably, dCKA shows a stronger correlation with domain similarity than CKA (Publication III)

using:

$$\hat{\alpha}_f^m = (\text{vec}(K^0)^T \text{vec}(K^0))^{-1} \text{vec}(K^0)^T \text{vec}(K_f^m). \tag{5.3}$$

After the deconfounded similarity structures are obtained with Eq.5.1, we use the same similarity measure to calculate the deconfounded representation similarity:

$$ds_{f_1,f_2}^{m_1,m_2} = s(dK_{f_1}^{m_1}, dK_{f_2}^{m_2}). \tag{5.4}$$

This deconfounding method improves reliability of neural network similarity metrics by ensuring they accurately reflect the true functional similarities between models, rather than being skewed by the structural traits of the input data. This technique is particularly valuable in transfer learning, where it's crucial to identify functional similarities between models trained on different domains.

The source code is available online.

5.2 Key Results

In Figure 5.1, the study demonstrates the application of deconfounded Centered Kernel Alignment (dCKA) on two pairs of ResNet-18 models – one altered by random Gaussian noise and another that is a pretrained model finetuned on CIFAR-10. Initial CKA analysis revealed surprisingly high similarity scores (0.99) between the networks altered randomly, pointing to significant confounding from input similarities. After applying covariate adjustment with dCKA, the similarity score for these models significantly decreased to 0.43. This experiment highlights the importance of deconfounding in ensuring similarity metrics accurately reflect true functional similarities, rather than being influenced by the structure of the input data.

dCKA better aligns with domain similarities

We study how the similarities computed using conventional CKA compares with dCKA during transfer learning under domain shift. We examine models that start from a pretrained (PT) base and are finetuned (FT) on data from different domains. We expect that the similarity among the FT models will reflect the natural similarities between their respective input domains.

To study this in detail, we utilize the Multi-lingual STS-B dataset [65], focusing on English, Spanish, Polish, and Russian as the target domains. We start with a pretrained (PT) XLM-RoBERTa model [13] and calculate layer-wise CKA and dCKA between each finetuned (FT) model and the PT model using the test set from the respective target domain [69]. To determine the similarity between two domains, we construct a cross-domain classifier with an equal number of samples from each domain [75]. Finally, we assess the correlation between the domain similarities (defined by the BCE loss from the cross-domain classifier) and the CKA and dCKA scores between the PT and FT models for each domain, using Spearman's ρ and Kendall's τ correlation [18].

We observe that dCKA correlates more closely with domain similarity than CKA. Figure 5.2 (*Left*) displays the results from the test BCE loss when a cross-domain classifier was trained to distinguish between domains on the STS-B dataset. The comparisons of layer-wise CKA and dCKA between the pretrained and finetuned XLM-RoBERTa models are depicted in Figure 5.2 (*Center* and *Right*), respectively. Models finetuned on highly dissimilar domains, such as English and Russian, tend to show lower layer-wise similarities with the pretrained model.

Limitations and Challenges

Despite the advancements presented, our approach faces a few limitations. For example, we assumed that the confounder is linearly separable from the representation similarity, but it remains possible that this separability is not fully additive, particularly in deeper layers. This may explain why deconfounding similarities appear more beneficial for shallow layers. One potential solution is to regress out the similarity structure in the previous layer rather than the input layer. However, this approach discards information from earlier layers, ultimately reducing the ability to represent similarities in functional behaviors. Addressing these issues and developing more functionally consistent similarity measures remains an open question for future research.

5.3 Revisiting Research Question 3

Research Question 3 addressed the challenge of more accurately aligning the representation similarity of neural networks with their functional similarity. Our research demonstrates that traditional metrics like CKA and RSA are influenced by the structural traits of input data, potentially skewing their assessments of model similarity. By implementing covariate adjustments, we have improved these metrics, particularly deconfounded CKA, which now exhibits a stronger correlation with true functional similarities. This adjustment effectively differentiates models finetuned on varied domains and aligns more accurately with actual domain similarities. Overall, our approach enhances the effectiveness of similarity metrics in assessing the success of transfer learning and contributes to the development of more interpretable and accurate machine learning models.

6. Contrastive Language-Image Training with Expert Annotations

Electronic Health Records (EHR) inherently feature multi-modal data like free-text medical notes from health professionals, imaging scans, and structured entries like lab results, drug prescriptions and diagnoses. By leveraging these diverse data types, machine learning models can provide a more comprehensive understanding of a patient's health dynamics. Recent advances in deep learning have introduced models capable of handling such multi-modality effectively. A notable example is CLIP (Contrastive Language-Image Pretraining), which excels at integrating textual and visual data. CLIP's primary mechanism involves mapping embeddings from different modalities into a shared space, thereby facilitating direct comparisons and improving interoperability.

However, recent studies have revealed a persistent challenge known as the "modality gap" [59]. This occurs when embeddings from different modalities form distinct clusters within the shared space, leading to a phenomenon known as cone effect. This effect can lead to mismatches and interference among non-paired embeddings. This issue is especially pronounced in the EHR domain, where the variability of data complexity and quality pose additional problems for multi-modal models [71].

Research Question 4: Can high-quality expert annotations improve the baseline performance of a medical CLIP model?

To tackle this, Publication IV explores the benefit of integrating high-quality expert annotations – specifically radiologists' eye-gaze heatmaps [5, 45] – into the CLIP training process. These heatmaps reveal the focal points of radiologists during X-ray evaluations, providing detailed annotations. We hypothesize that these granular insights can enhance the model's performance within a contrastive learning framework.

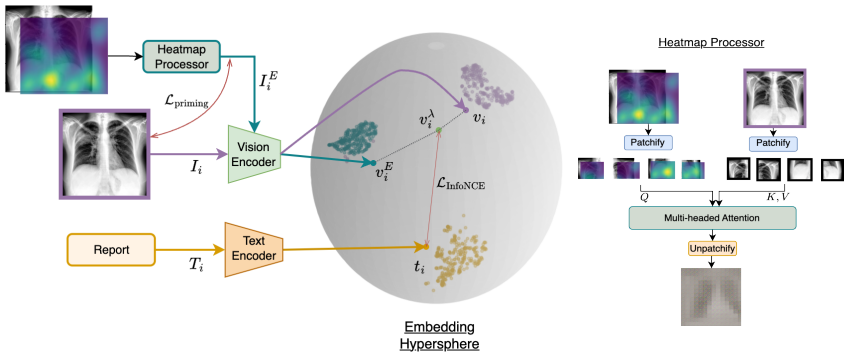

Figure 6.1. eCLIP Pretraining with Expert Annotations. eCLIP introduces a Heatmap Processor *(right)*, equipped with a multi-headed attention layer, to the usual Image and Text encoders used in CLIP. The original image (I_i), its corresponding text (T_i), and the heatmap-processed image (I_i^E) are presented within a tripartite area. We utilize a mixup technique between I_i and I_i^E to create the embedding v_i^λ, facilitating additional positive pairs to improve the optimization of the CLIP InfoNCE loss. Additionally, an auxiliary loss, $\mathcal{L}$priming, is employed during initial training steps to 'prime' the heatmap processor to output the original image when the heatmap consists entirely of ones (Publication IV)

6.1 Method

Contrastive learning models like CLIP often face challenges with high intra-modal similarity, where similar types of images, such as two chest X-rays showing different pathologies, might appear closely related. This closeness can inflate the perceived similarity among negative pairs in the InfoNCE loss function (Eq. 2.9), leading to a decrease in the model's ability to differentiate between them effectively. This issue is commonly addressed by introducing hard negative pairs, such as by mixing embeddings from different modalities [71], which improves discrimination but may reduce the semantic clarity of the embeddings. Alternatively, our approach uses expert-annotated heatmaps overlaid on images to generate additional positive and negative pairs (see Figure: 6.2). These heatmap-enhanced embeddings (I^E) form a new, distinct cluster in the shared embedding space, as illustrated in Figure: 6.1 (Left).

This section outlines the components of the expert-annotated CLIP model (eCLIP), which includes a heatmap processor for generating I^E embeddings, mixup augmentation to balance the disparity between the scarce expert annotations and the larger training dataset, and a curriculum learning strategy [3] to smoothly introduce expert annotations without disrupting the initial training dynamics.

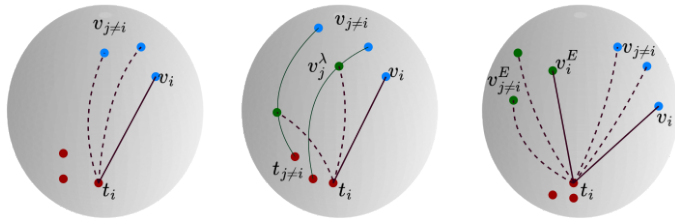

Figure 6.2. Comparing eCLIP with m^2-mixup[71]. *(left)* Standard CLIP showing image-text positive pairs (v_i, t_i) (solid line), while the other image embeddings serve as negative pairs (dashed line). *(center)* the m^2-mixup creates negative pairs (v_j^{λ}, t_i) via interpolation between embeddings along the geodesic. *(right)* eCLIP adds expert image embedding, v_i^E, in addition to v_i for text t_i, forming additional positive and negative pairs (Publication IV)

6.1.1 Heatmap Processor

The heatmap processor transforms the images and their corresponding heatmaps into sequences of 16×16 patches. It applies multi-headed attention, treating the patched images overlaid with heatmaps as queries and the original image patches as keys and values (Figure: 6.1, Right). This processing step integrates expert insights directly into the image embeddings, which is subsequently used to generate positive samples for contrastive pretraining of CLIP.

To ensure robust performance in the absence of expert annotations, a **priming** mechanism is introduced. This involved training the heatmap processor to output the original image when the expert annotations are not available, via a mean-squared error loss, $\mathscr{L}_{\text{priming}}$. This ensures that the model remains effective even when the expert data is unavailable.

6.1.2 Mixup Augmentation

Given the limited availability of high-quality expert annotations – approximately 1,000 annotations within a dataset of around 200,000 image-text pairs – eCLIP employs mixup augmentation to effectively utilize these valuable yet scarce resources. Mixup augmentation involves blending images annotated by experts with those that are not. The proportion of blending is determined by a variable $\lambda \sim \text{Beta}(\alpha = 0.3, \beta = 0.3)$. This technique not only increases the diversity of the training samples but also bolsters the model's ability to generalize from a limited set of expert data. By interpolating between images, the model learns to smooth out the decision boundaries, which can reduce overfitting and improve the robustness of the predictions. In practice, this means that eCLIP is better equipped to handle real-world variations in medical images, leveraging expert insights effectively even when such annotations are sparse.

6.1.3 Curriculum Learning Strategy

To ensure smooth integration of expert annotations without disrupting the foundational training, we implement a phased curriculum learning strategy that allows for gradual adaptation to the complexity of the annotated data. The strategy begins with a **Cold Start Phase**, where the model initially trains solely on non-annotated data to establish a robust baseline, accounting for approximately 10% of total training iterations. This is followed by the **Warmup Phase**, during which the inclusion of expert-annotated examples gradually increases from a probability of 0.05 to 0.5 over the next 30% of training iterations, enabling the model to progressively adjust to the new data. Finally, the **Cooldown Phase** reduces the probability of using expert annotations to 0.1 for the remaining 40% of the training, fine-tuning the model's response to expert-driven insights and balancing it with the foundational training.

This structured approach optimizes the model's performance, ensuring a stable integration of expert knowledge while maintaining generalizability. The source code is available online.

6.2 Key Results

Our experiments were structured to assess the impact of expert heatmap annotations on the quality of the learned embeddings in the eCLIP model. These experiments utilized a significant number of image-text pairs, a small fraction of which were annotated with eye-gaze heatmaps, to train the model. Importantly, no annotations were used for the test samples, ensuring that the model's generalization capabilities were rigorously evaluated.

Adding expert annotations improves sample efficiency
We study the model's sample efficiency by limiting the amount of training data that is available for training. Our findings indicate that eCLIP with expert annotations enhances the sample efficiency, outperforming the baseline CLIP model across multiple datasets. Specifically, in zero-shot classification performance tests – which adjusted the number of training batches available for pretraining – eCLIP demonstrated superior efficiency on MIMIC 5x200 [42, 44], CheXpert 5x200 [39] and CXR 24x100 [96] datasets (Figure 6.3, Top).

Further analysis using a linear probe three class-imbalanced datasets – CXR-8, RSNA and OpenI-5, revealed that eCLIP also shows stronger performance in multi-label classification tasks. For example, it excelled in the CXR-8 [96] and OpenI-5 [15] datasets and remained competitive in binary classification for RSNA [86] as seen in Figure 6.3, Bottom.

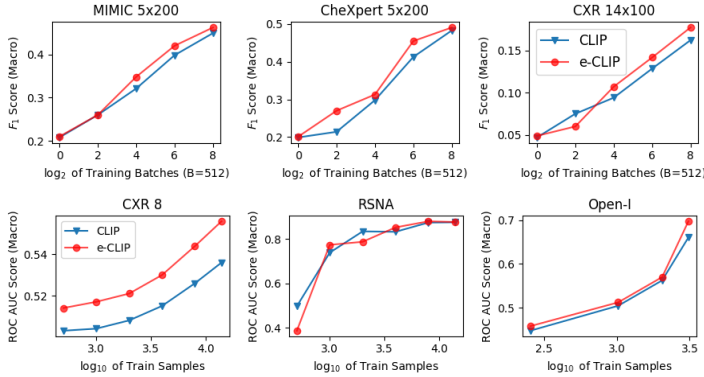

Figure 6.3. Sample Efficiency. *(Top row)* Compares zero-shot performance for CLIP and eCLIP Swin Tiny models on three multi-label classification test sets, trained over varying number of training batches. *Bottom row* Shows linear probe scores achieved with varying amounts of training data (Publication IV)

Limitations and Challenges

This study is constrained by the limited size of expert-annotated data, which prevents a comprehensive analysis of the impact of annotation size or distribution across different abnormalities. Additionally, eCLIP incurs higher computational costs during training due to the additional forward pass required in the warmup and cool-down phases. Furthermore, the clinical relevance of the generated radiology reports has not been validated by medical professionals, relying instead on standard metrics that may introduce biases and fail to fully capture clinical accuracy [102].

6.3 Revisiting Research Question 4

Research Question 4 inquired, *"Can high-quality expert annotations improve the baseline performance of a medical CLIP model?"* The integration of expert eye-gaze heatmaps into the eCLIP model significantly enhanced sample efficiency and improved the model's generalization across multiple datasets. These findings confirm the benefits of leveraging high-quality, albeit limited, expert data to enhance model performance. Specifically, by effectively transferring knowledge between expert-annotated and non-annotated data, eCLIP demonstrates the ability to learn from limited data inputs. This capability to improve learning efficiency and model accuracy through expert annotations promises to enhance diagnostic accuracy and patient outcomes, paving the way for more personalized and accurate healthcare solutions.

7. Conclusion

Through a systematic exploration of four distinct research questions, this thesis demonstrates the benefits of adapting advanced machine learning techniques to Electronic Health Records (EHR), significantly enhancing both prediction accuracy and operational efficiency.

Firstly, the study confirms that integrating temporal dynamics through sequential models like RNNs improves healthcare utilization forecasts. This highlights the critical role of accounting for the temporal aspect of patient histories in predicting future healthcare needs more effectively.

Secondly, the development of specialized models tailored to EHR data proved superior to generic NLP models, reinforcing the value of custom solutions in healthcare analytics. These specialized models leverage the unique structure and patterns of EHR data, thus providing more accurate and efficient predictions.

Thirdly, by adjusting for confounders in traditional similarity metrics like CKA and RSA, the research has shown that it is possible to obtain a more accurate reflection of functional similarities between neural networks. This enhancement is particularly crucial for applications such as transfer learning, where understanding the functional capabilities of models across domains is key.

Finally, the exploration of multi-modal contrastive learning with expert annotations within a medical CLIP model illustrated that high-quality, expert-driven annotations could significantly refine model performance. This approach improves the model's ability to correlate and interpret complex visual and textual medical data, leading to better diagnostic tools.

While our research demonstrates the efficacy of deep learning in addressing a few complex problems in healthcare, it still confronts significant challenges, particularly around interpretability. Our efforts to enhance interpretability through comparative methods like CKA and dCKA provide some insight, yet understanding the underlying reasoning of model's decisions remains elusive. This is particularly critical in healthcare, where both accuracy and clarity in decision-making processes are equally important. Although attention mechanisms used in sequential models are

believed to enhance transparency by highlighting influential parts of the input, recent research suggests that the attention weights do not often correlate with the actual contributions of input features [40]. Moving forward, Bayesian neural networks [64, 25] present a promising avenue for improving model transparency by quantifying prediction uncertainty. This approach could allow for human intervention when model confidence is low, thereby enhancing reliability in critical applications. However, these models do not scale well to larger datasets or more complex model architectures. Developing techniques that bridge the gap between Bayesian methods and large-scale applications could be a highly promising research direction.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, 2009.

[4] Bent Guttorm Bentsen. International classification of primary care. *Scandinavian Journal of Primary Healthcare*, 4(1):43–50, 1986.

[5] Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific Data*, 9(1):350, 2022.

[6] Sabrina Bonomi. The Electronic Health Record: A comparison of some European countries. In *Information and Communication Technologies in Organizations and Society: Past, Present and Future Issues*, pages 33–50. Springer, 2016.

[7] Gerlind R Brämer. International statistical classification of diseases and related health problems. Tenth revision. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, 41(1):32–36, 1988.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[11] Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, pages 3512–3520, 2016.

[12] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[14] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[15] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing. *arXiv preprint arXiv:2108.01661*, 2021.

[19] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[20] Randall P Ellis, Bruno Martins, and Sherri Rose. Risk adjustment for health plan payment. In *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets*, pages 55–104. Elsevier, 2018.

[21] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[22] EU Healthcare. Medical Records in Finland. Accessed: 27-april-2024.

[23] R Scott Evans. Electronic Health Records: Then, Now, and in the Future. *Yearbook of Medical Informatics*, 25(S 01):S48–S61, 2016.

[24] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[25] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference onMmachine Learning*, pages 1050–1059. PMLR, 2016.

[26] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[27] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[29] Sebastien Haneuse, David Arterburn, and Michael J Daniels. Assessing missing data assumptions in ehr-based studies: a complex and underappreciated task. *JAMA Network Open*, 4(2):e210184–e210184, 2021.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[32] Tianxing He, Shengcheng Yu, Ziyuan Wang, Jieqiong Li, and Zhenyu Chen. From data quality to model quality: An exploratory study on deep learning. In *Proceedings of the 11th Asia-Pacific Symposium on Internetware*, pages 1–6, 2019.

[33] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

[34] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[35] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[36] Tuukka Holster, Shaoxiong Ji, and Pekka Marttinen. Risk adjustment for regional healthcare funding allocations with ensemble methods: an empirical study and interpretation. *The European Journal of Health Economics*, pages 1–15, 2024.

[37] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[38] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. pmlr, 2015.

[39] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[40] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556, 2019.

[41] Ali Jazayeri, Ou Stella Liang, and Christopher C Yang. Imputation of missing data in electronic health records based on patients' similarities. *Journal of Healthcare Informatics Research*, 4(3):295–307, 2020.

[42] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG-Chest radiographs with structured labels. *PhysioNet*, 2019.

[43] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[44] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[45] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, 8(1):92, 2021.

[46] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 2017.

[47] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69, 2022.

[48] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*, 33:4163–4174, 2020.

[49] Young-Gun Kim, Kyoungwon Jung, Young-Taek Park, Dahye Shin, Soo Yeon Cho, Dukyong Yoon, and Rae Woong Park. Rate of electronic health record adoption in South Korea: a nation-wide survey. *International Journal of Medical Informatics*, 101:100–107, 2017.

[50] Rohan Kodialam, Rebecca Boiarsky, Justin Lim, Aditya Sai, Neil Dixit, and David Sontag. Deep contextual clinical prediction with reverse distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 249–258, 2021.

[51] Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34, 2021.

[52] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[53] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[54] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[55] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

[57] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with Fourier Transforms. *arXiv preprint arXiv:2105.03824*, 2021.

[58] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):1–12, 2020.

[59] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multimodal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[60] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[61] Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. Pay attention to MLPs. *Advances in Neural Information Processing Systems*, 2021.

[62] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[63] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[64] David John Cameron Mackay. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

[65] Philip May. Machine translated multilingual STS benchmark dataset. 2021.

[66] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(1):1–12, 2020.

[67] Luke Melas-Kyriazi. Do you even need attention? A stack of feed-forward layers does surprisingly well on Imagenet. *arXiv preprint arXiv:2105.02723*, 2021.

[68] Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129, 2021.

[69] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *arXiv preprint arXiv:2008.11687*, 2020.

[70] Office of the National Coordinator for Health Information Technology. What information does an electronic health record (ehr) contain?, 2019. Accessed: 27-april-2024.

[71] Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *Advances in Neural Information Processing Systems*, 36, 2023.

[72] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[73] Mary Phuong and Marcus Hutter. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*, 2022.

[74] Konstantinos Psychogyios, Loukas Ilias, Christos Ntanos, and Dimitris Askounis. Missing value imputation methods for electronic health records. *IEEE Access*, 11:21562–21574, 2023.

[75] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018.

[76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[77] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[78] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[79] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):1–13, 2021.

[80] Sherri Rose. A machine learning framework for plan payment risk adjustment. *Health Services Research*, 51(6):2358–2374, 2016.

[81] Alexander M Rush. The annotated transformer. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 52–60, 2018.

[82] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47. IEEE, 2013.

[83] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[84] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[85] Mahdiyar Shahbazi, Ali Shirali, Hamid Aghajan, and Hamed Nili. Using distance on the Riemannian manifold to compare representations in brain and in models. *NeuroImage*, page 118271, 2021.

[86] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[87] Akritee Shrestha, Savannah Bergquist, Ellen Montz, and Sherri Rose. Mental health risk adjustment with clinical categories and machine learning. *Health Services Research*, 53:3189–3206, 2018.

[88] Sunil Kumar Srivastava. Adoption of electronic health records: a roadmap for India. *Healthcare Informatics Research*, 22(4):261, 2016.

[89] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. Data processing and text mining technologies on electronic medical records: a review. *Journal of Healthcare Engineering*, 2018(1):4302425, 2018.

[90] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. MLP-Mixer: An all-MLP architecture for vision. *Advances in Neural Information Processing Systems*, 2021.

[91] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

[92] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[93] Tuulikki Vehko, Salla Ruotsalainen, and H Hyppönen. E-health and e-welfare of Finland. *Check point*, 2019, 2018.

[94] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[95] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[96] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2097–2106, 2017.

[97] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[98] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[99] Karen Wilson and Lara Khansa. Migrating to electronic health record systems: a comparative study between the United States and the United Kingdom. *Health Policy*, 122(11):1232–1239, 2018.

[100] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

[101] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. CXR-CLIP: Toward large scale chest X-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023.

[102] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

[103] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[104] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[105] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[106] Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. CLIP in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

Aalto-DT 197/2024

ISBN 978-952-64-2030-1 (printed)
ISBN 978-952-64-2031-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES