Simplifying ICD-10 Coding: Evaluating Pre-trained Language Models and Traditional Methods for Clinical Notes Mapping

This Research Project report is submitted to the Department of Computer Science as partial fulfillment of Master of Science in Computer/Data Science degree.

by

Syed Bilal Rizwan

Supervised by
Prof. Dr. Sajjad Haider
Professor
Department of Computer Science
School of Mathematics and Computer Science (SMCS)
Institute of Business Administration (IBA), Karachi

Fall Semester 2023
Institute of Business Administration (IBA), Karachi, Pakistan

Simplifying ICD-10 Coding: Evaluating Pre-trained Language Models and Traditional Methods for Clinical Notes Mapping

This Research Project report is submitted to the Department of Computer Science as partial fulfillment of Master of Science in Computer/Data Science degree.

by

Syed Bilal Rizwan
(ERP ID: 23943)

Dr. First name Last name _____ Sajjad Haider_____
Supervisor
University Affiliation

Dr. First name Last name ___ Shakeel A. Khoja____
Dean
School of Mathematics and Computer Science (SMCS)
Institute of Business Administration (IBA), Karachi

Fall Semester 2023

Institute of Business Administration (IBA), Karachi, Pakistan

Dedication

This project is committed to the unwavering search for innovation and quality in the healthcare technology sector. This work is a tribute to all the people who work so hard to close the gap between medical knowledge and state-of-the-art computational solutions.

This project is dedicated to patients whose well-being is at the heart of every medical advancement. Your stories motivate us to strive further to push the boundaries of what is possible.

Furthermore, this project is dedicated to my mentor, advisor, and supervisor Prof.Dr Sajjad Haider who provided guidance and support throughout the project, your wisdom and support has been invaluable.

Lastly, I would like to extend my dedications to my family and friends whose unwavering support has been the bedrock of my journey.

I hope that this initiative will stand as a testament to the passion, hard work, and dedication of everyone who contributes to the development of technology and healthcare.

Acknowledgement

I extend my heartfelt gratitude to all those who contributed to the realization of this project. My sincere thanks go to Dr Sajjad Haider for being a valuable mentor and supervisor who guided me throughout the project. Your insights have been instrumental in shaping the direction of our work.
Then, to PhysioNet in giving access and letting me utilize their rich MIMIC IV with coded labels. The dataset helped in the development of the methodology and testing.
Furthermore, the transformer-based pre-trained models available to use by Emran [2] and embedding models such as PubMedBERT helped greatly in testing different methodologies to come up with the best product.
A special acknowledgment is reserved for our families and friends whose encouragement and understanding sustained us during the challenges of this project.
The project stands as a collective endeavor and I am grateful to everyone who played a role, no matter how small, in bringing it to fruition.

Table of Contents

Abstract

The project explores the challenges and importance of medical coding in healthcare industry focusing on the International Classification of Diseases, Tenth Revision (ICD-10) coding system. The goal of this project is to do a comparative analysis between pre-trained language model in ICD-10 assignments to clinical notes. The dataset employed in this project is MIMIC IV, and most methodologies are tested on the shortest 1,000 clinical notes from the MIMIC IV dataset, predicting the first three letters of the ICD-10 code (category). Several methodologies are experimented with in this project, benchmarked against the performance of MEDCAT's SNOMED-CT. It was observed that SNOMED outperformed all methodologies with a Macro-F1 of 0.218, whereas the best transformer-based methodology yielded a Macro-F1 of 0.123. This result is not considered low, given the substantial number of classes being predicted, which is around 24,000 for overall MIMIC IV dataset and 700 for our testing dataset. Furthermore, in comparison with related works, both best approaches outperformed metrics from the research papers.

The comparative analysis paves a way for finding the best approach to map clinical notes to ICD-10 codes which can handle various sizes and styles of texts.

Keywords:

Transformers, BERT, GPT, Deep Learning, Medical Coding, ICD10, MEDCAT, SNOMED-CT, Natural Language Processing, Text Cleaning, NER.

Chapter 1

1.1 Background and Context

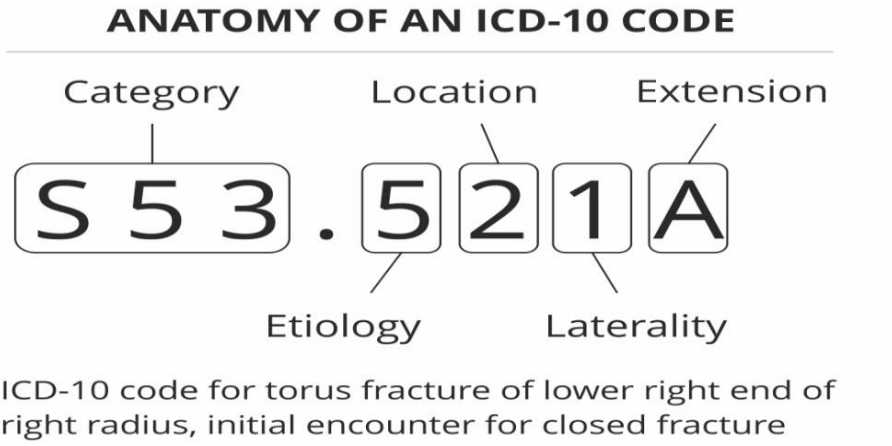
The healthcare industry is dynamic and data-driven, evolving constantly to improve patient care, operational efficiency, and service quality. The usage of standardized coding systems dates to the 1800s, when the importance of such systems became clear due to nomenclature irregularities and the need for better statistical data. The International Statistical Institute published the first international classification of illnesses (ICD) in 1893. This classification system was quickly adopted in the US, Canada, and Mexico. It underwent considerable adjustments over the years under the auspices of the World Health Organization (WHO), eventually becoming a crucial tool for healthcare professionals, academics, and healthcare billing systems all around the world.

1.1.1 What Is ICD-10 Codes?

The International Classification of Diseases, Tenth Revision (ICD-10) is a worldwide recognized and standardized system of alphanumeric numbers used to identify diseases, disorders, and other health-related issues. The World Health Organization (WHO) developed and maintains ICD-10, which provides a consistent language for healthcare providers to document, monitor, and communicate medical conditions and diagnoses.

Each ICD-10 code denotes a unique medical diagnosis, symptom, or procedure. These codes cover a wide range of medical illnesses, from common maladies to unusual diseases. They allow for the consistent recording and retrieval of medical data, which is critical in record-keeping, billing systems, and epidemiological research. Because it allows for precise data management, this standardized coding system is an essential tool for healthcare professionals, insurance companies, healthcare organizations, and academics. Here is what an ICD-10 code looks like:

1.1.2 How Do They Look Like?



1.1.3 Challenges with Medical Coding

Following the worldwide introduction of the ICD-10 system, healthcare practitioners' manual coding of patient diagnoses to ICD-10 codes has become a time-consuming and

error-prone operation. ICD-10 has many codes, over 73,000 in total, that cover a broad range of diseases, procedures, and medical conditions. Because of its complexity, hand coding is extremely intricate and time-consuming. As a result, hospitals are frequently forced to engage more personnel or interns to undertake this work, which might unfortunately lead to an increase in errors.

Furthermore, healthcare is a dynamic industry in which new diseases and procedures occur on a regular basis, needing continuous revisions to the coding system. Medical coders must stay up to date on these changes, which adds to the complexity of their responsibilities. In addition, some areas of coding necessitate a thorough understanding of medical language, anatomy, and disease classifications. This demand for highly skilled and specialized personnel in medical coding adds to the task's mental and financial resource intensity for healthcare organizations.

1.1.4 Importance of Medical Coding

Before addressing the challenges that medical coding presents, it is critical to understand its significance. Medical coding accuracy ensures that patient records are complete and precise, providing healthcare professionals with a thorough understanding of a patient's medical history. This comprehensive understanding enables practitioners to provide the best possible patient care.

Moreover, medical coding provides hospitals with the tools they need to perform accurate billing and secure timely reimbursement for the healthcare services they provide. It also assists legal and compliance teams in ensuring regulatory standards, laws, and regulations are followed. Finally, medical coding assists healthcare researchers and statisticians in carrying out high-quality epidemiological studies, monitoring public health, and furthering medical research.

However, given the formidable challenges that this task entails, it becomes imperative to develop efficient automated or semi-automated solutions for clinical coding.

1.2 Project Objective

The primary goal of this project is to try different methodologies and do a comparative study to address the challenges associated with mapping clinical notes to ICD-10 codes. To ensure accurate and timely mapping of clinical notes to ICD-10 codes, this study will try transformer-based approaches, with a particular emphasis on BERT (Bidirectional Encoder Representations from Transformers) and natural language processing (NLP) techniques. To assess its efficacy, the performance of these approaches will be compared to that of existing methods, such as the use of SNOMED CT and the MEDCAT library.

1.3 Dataset Used

The MIMIC-IV dataset was used in this project; it is a large and impressive collection that includes records for 180,733 different patients and 431,231 hospital admissions. This dataset contains discharge summaries for each patient, as well as the ability to map these summaries to corresponding ICD-10 codes, which are linked to each admission_id (hadm_id), resulting in labelled codes for each discharge summary.

1.4 Scope and Limitations

This project's scope is purposefully limited to transformer-based approaches. For benchmarking purposes, SNOMED-CT and MedCat are used. The project scope specifically includes the prediction of only the first three digits of the ICD-10 code. This limitation aims to reduce the number of codes used while still providing useful information about the diagnosis category. The first three digits represent the primary diagnosis category, which is both informative and useful information. Furthermore, only diagnosis codes will be predicted. Adding procedure ICD-10 codes will further increase the complexity of the task making it difficult to achieve acceptable metrics.
The transformer-based approaches will primarily include the use of the GPT-3.5 API, prompt engineering, NER-tagging, and BERT classification techniques to complete the task.

1.5 Key Insights

In this project, various methodologies were attempted; however, the utilization of MEDCAT's SNOMED-CT model-pack yielded the best results. This is likely due to the fact that the MEDCAT library employs advanced natural language processing (NLP) and deep learning techniques to extract SNOMED-CT concepts, which can then be easily mapped to ICD-10 codes using a simple dictionary mapper. The transformer pre-trained language models tested could not surpass MEDCAT for this specific reason. Nevertheless, they do serve as a valuable foundation for addressing the problem using transformer-based models.

While experimenting with the GPT methodology, it became apparent that large language models (LLMs) like GPT/BARD are still not sufficiently equipped to handle the intricacies of ICD-10 coding.

Furthermore, promising results were observed with pre-trained transformer-based models such as Emran's BERT. However, these models were fine-tuned on very short clinical summaries, primarily resulting in inferior performance on comprehensive hospital clinical notes like those in MIMIC IV. Overcoming this challenge prompted the development of a rigorous preprocessing technique that significantly enhanced the performance of every model, including MEDCAT. The transformer-based models were provided with extracted sentences from the clinical notes after preprocessing, enabling them to classify ICD-10 code categories more effectively.

There is considerable potential for achieving superior metrics with language models such as BERT/ALBERT if sufficient resources and time are allocated for fine-tuning them on MIMIC-IV clinical notes, which are lengthy and information-rich.

Finally, the evaluation metrics employed—Macro-F1, Precision, and Recall—were primarily chosen because they were the most used metrics in relevant literature for performance comparison. However, it is worth noting that Micro metrics could have proven useful in this context as well. Given the high-class imbalance, where some ICD-10 codes appear frequently (as expected for common diseases like fever) and others are rare

(as seen with diagnoses like Keratoconus), a model detecting Keratoconus well but not performing as effectively for fever could yield a higher Macro-F1 than Micro.

1.6 Report Organization

The report is organized as follows this section: The second chapter provides a thorough review of the literature on ICD-10 coding of clinical notes, involving SNOMED-CT, Medcat library, and LLM classification. Chapter 3 presents the methodologies tried in the project, while Chapter 4 focuses on results obtained from those experiments. Chapter 5 discusses in detail the findings from the experiments and comparison with literature review and report and concludes with chapter 6 as conclusion.

Chapter 2: Technical Background and Literature Review

2.1 Technical Background

This project entails the primary tasks of disease category code detection from clinical notes, specifically the first three digits of ICD-10 codes. The methodologies currently in use and those that will be investigated are described in detail in the following sections. It is critical to note that all clinical notes will be cleaned using NLP techniques.

2.1.1: Utilizing MEDCAT with SNOMED-CT Concepts:

This method will serve as a metric benchmark, allowing for performance comparisons with other methods. It is especially valuable because it is a methodology that is currently being used by several systems. A SNOMED-CT model pack will be imported in this approach to map ICD-10 codes to each processed clinical note. Following that, metrics such as Macro F1, Precision, and Recall will be calculated and saved for comparison with other methods under consideration.

2.1.2: Leveraging Prompt Engineering With GPT:

Each ICD-10 diagnosis code is accompanied by a description. This method aims to capitalize on these descriptions by generating Cohere embeddings for each ICD-10 description and storing them in a vector database like Pinecone. Using the Cohere model, the embedding for a clean clinical note is computed. Following that, based on cosine similarity, the top 50 descriptions most closely related to the clinical note will be retrieved from the vector database. The next step is to use these 50 ICD-10 codes, along with their descriptions, to test various prompts. The context, along with the clinical note, will be sent to the GPT API via a carefully designed prompt, with the goal of mapping the clinical note to the corresponding ICD-10 codes. However, comparing clinical note embeddings, which typically exceed 10,000 characters, with ICD-10 description embeddings, which typically contain fewer than 200 characters, presents a significant technical challenge.

2.1.3: Harnessing Pre-trained BERT:

This method entails importing various pre-trained BERT models from the literature. These models are either pre-trained on medical data or custom-built for the task at hand: categorizing clinical notes into ICD codes or categories. Several types of BERT are available online, and their results will be compared.

2.1.4: Using PubMedBERT Embeddings:

This method involves the use of PubMedBERT embedding model to compare embeddings of clinical note with ICD-10 descriptions. The idea is to use a vector database such as pinecone to store ICD-10 category descriptions and then utilize a vector search against the note’s embedding to find the closest ICD-10 codes.

2.1.5: Using ALBERT-NER with PubMedBERT Embeddings:

The idea is like section 2.1.4 however, there's an additional step in between where NER model Albert-medical-ner is used to extract entities from a clinical note and their embeddings are then compared with pre-stored ICD-10 category embeddings.

2.2: Literature Review

This section is divided into subsections, each devoted to research methodology. While the GPT methodology is new, making relevant research papers scarce, a thorough analysis of research papers related to the well-established SNOMED-CT and BERT methodologies is presented below.

2.2.1: Clinical Coding Using SNOMED-CT Concepts

The MEDCAT library stands out as a resource with exceptional tools for extracting ICD-10 codes and SNOMED-CT concepts, and SNOMED-CT concepts provide a direct avenue for mapping to ICD-10 codes. The procedure begins by mapping clinical notes to SNOMED-CT concepts, which are then converted to ICD-10 codes.

One relevant research paper, "Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based Approach Using SNOMED CT to ICD-10 Mappings," investigates this methodology. The primary goal of this paper is to introduce a natural language processing (NLP) approach for hospital in-patient coding that leverages mappings between SNOMED CT and ICD-10-AM (Australian Modification). This research is an early attempt to automate the clinical coding process. The study's dataset includes 569,846 patient encounters from three Australian hospitals within the Gold Coast Hospital and Health Service (GCHHS) from January 2011 to December 2015. Notably, the paper uses only NLP techniques to map SNOMED-CT to ICD-10, with no advanced machine learning models involved. The sensitivity and positive predictive value (PPV) of the proposed NLP-based approach that uses SNOMED CT for diagnostic coding are reported as evaluation metrics. The study ends with a sensitivity rate of 54.1% and a PPV of 70.2%. It is worth noting that the method's simplicity contributes to the low metrics reported in this work. [1]

2.2.2: Clinical Coding Using BERT

One notable research paper on the application of BERT stands out as particularly promising. This paper, titled "Improving clinical documentation: automatic inference of ICD-10 codes from patient notes using a BERT model," makes a few important contributions. It presents a novel Pipeline BERT approach for categorizing clinical notes into ICD-10 codes. The study also compares the efficacy of the BERT methodology to baseline models based on LSTM and CNN. The study made use of a dataset obtained from the KAUH health Centre, which included an examination of 7,541 clinical notes. This dataset contains a total of 15,747 distinct classes for full ICD-10 codes, divided into 1,701 distinct categories. Several deep learning models, including LSTM and GRU, were investigated in the study as baseline models. Not only was BERT used in the experiment, but so was a pipeline of BERT models. This pipeline approach, in which the first BERT predicts the category and the second predicts the rest of the code, produced positive results. The Macro F1-score was used as the evaluation metric in this paper, with the pipeline

model achieving an impressive F1 Score of 92.5% based on a test set of 1,500 clinical notes. It is important to note, however, that one major limitation of this work is the relatively small size of the dataset used, which does not include all ICD-10 codes. [2]
Another paper of interest, titled "MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT," demonstrates the effectiveness of transfer learning using the pre-trained language representation model, BERT, and its variant, BioBERT. The experiment used a dataset of 8,385 training documents, including a development set, and 407 test documents, all of which were originally written in German and later translated into English. The study compared their experimental approach with BERT to various baseline models, including SVM, CNN, Stacked LSTM, Conditional LSTM, and an attention-based model. The performance metric of choice was F1-micro, and the best model, BioBERT, achieved a remarkable F1-micro score of 73% on the testing set, along with 86% recall and 64% precision. However, it is important to note that this study had limitations, which were primarily related to the focus on a single domain, namely animal experimentation, and the relatively small size of the testing set. [3]
A third paper, "PLM-ICD: Automatic ICD Coding with Pretrained Language Models," makes a significant contribution by introducing the PLM-ICD framework for automatic ICD-10 coding using pretrained language models. Using the benchmark MIMIC II and MIMIC III datasets, the experimental results show that their proposed framework achieves state-of-the-art performance across multiple metrics. MIMIC II contains 22,000 discharge summaries, whereas MIMIC III contains 52,000 discharge summaries. As the framework's backbone, the PLM model incorporates a Transformer architecture. The BioBERT model was specifically used by the authors on the MIMIC dataset. Several baseline models were developed and tested, including logistic regression, SVM, and neural networks. The performance metrics used included macro and micro F1, with the best macro-F1 and micro-F1 achieved using RoBERTa-PM at 10.4 and 59.8, respectively. However, it is critical to recognize that the limitations of this study are related to the models' relatively low metrics. [4]

2.3: Gap Analysis

While considerable progress has been made in automating the mapping of clinical notes to ICD-10, most of these achievements remain in the realm of research. The major gap in the research seems to be the size of the clinical notes used to train and test the models.
Furthermore, the landscape of discharge note data differs from hospital to hospital due to differences in how each institution records information. As a result, different preprocessing techniques will be needed to clean and prepare each clinical note for use as model input. For example, the data from a local Pakistani hospital differs significantly from the current benchmark dataset, MIMIC IV, which is used to compare model performance. As a result, dealing with these variations in data and discharge summary writing will be a unique challenge in this project.
The Imran and Co’s Paper on BERT classifier [2] showed a dummy example applying their model on ICD-code category descriptions. The major gap seen in the paper was that the texts given to the classifier were noticeably short since descriptions are less than 20 words whereas clinical notes are as long as 2000-10000 words. Furthermore, the model only gives a single ICD-code category as prediction. These two gaps from the paper will be addressed

in our project by devising methods that will allow bigger token lengths and give multiple ICD code categories as predictions.
Lastly, there are no advanced machine learning algorithms used in the SNOMED paper [1] which amounts to a lower performance. In this project we would aim to improve performance either by employing a smarter preprocessing technique or employing deep learning transformer-based techniques.

2.4: In Hindsight

Following initial GPT experimentation, it became clear that the methodology was not a viable option. The MedCAT library is the primary focus of this project's experiments, which use SNOMED-CT to map clinical notes to SNOMED concepts and then convert them to ICD-10 codes using a predefined map. After benchmarking and calculating metrics, the project's focus will shift to the pretrained language models, as discussed in Section 2.1.3.

Chapter 3: Methodology

3.1 Overview

Several methodologies were attempted in this project, beginning with the exploration of existing libraries. In this phase, the MEDCAT library and its SNOMED-CT model pack were imported and utilized to map processed clinical notes to ICD-10 categories. Initially, full ICD-10 codes were predicted, yielding unsatisfactory results. Consequently, the project scope was narrowed down to predicting ICD-10 categories (first 3 letters).

In the second phase, a GPT-based methodology was implemented. This involved creating embeddings of ICD code descriptions and saving them into a vector database called Pinecone. Subsequently, the embedding of a preprocessed clinical note was computed, and a vector similarity search identified the top-k ICD code descriptions that matched the clinical note based on cosine similarity. These top-k ICD code descriptions were then used as context for GPT, with the unprocessed clinical note provided as input for code prediction (One-Shot learning). Multiple prompts were experimented with in this section.

However, the second-phase methodology did not yield promising results, prompting a modification by excluding the GPT component. The revised approach involved creating embeddings for each ICD-10 category description using a pre-trained BERT model on the MIMIC dataset, specifically PubMedBERT (Yu Gu and Co, 2020 [10]). The processed clinical note was segmented into sentences, and the embedding for each sentence was generated using the same model. A vector search identified the top ICD-10 category description, which was then labeled accordingly. Thresholding was applied to the similarity score in the ultimate step to enhance the precision of the overall process.

The second last phase of the experiment involved using pre-trained BERT classifiers to multi-label classify ICD-10 categories for clinical notes. The model from the research paper by Emran and colleagues (Al-Bashabsheh and co [2]) was imported from Hugging Face and applied to the processed clinical note, which had been divided into sentences. Predicted categories from the BERT classifier model underwent thresholding on scores to improve the precision of the overall process.

The last phase of experiment involved a unique approach to extract disease entities using a transformer-based NER tagger named albert-medical-ner-proj from huggingface website. Meanwhile, category description embeddings are already saved in pinecone vector DB. After extracting entities, embeddings of those entities are created using PubMedBERT embedding model and a vector search is used to find the closest ICD-10 category for each entity. Lastly, a thresholding is applied on cosine similarity to weed out bogus matches to improve precision.

All these methodologies, once implemented on a small testing set of 1000 clinical notes, were evaluated based on various metrics, including Macro and Micro F1, Precision, Recall, and the Top 50 most occurring ICD codes. The processes and flowcharts for each methodology are available in section 3.3.

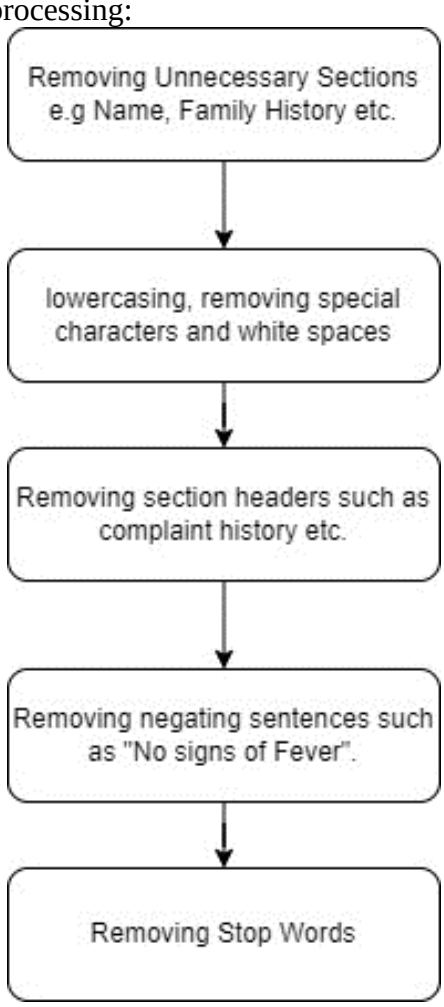
3.2 Importance

A significant distinction between existing research solutions and this project lies in the fact that this project also focuses on ways we can handle different sizes of texts with differing styles. Additionally, in comparison to other research solutions, this tool aims to provide multiple ICD-10 code categories for a single note, recognizing that each patient summary may involve multiple diagnoses/procedures. The incorporation of pretrained language models in this project seeks to apply a modern approach to the problem, addressing its challenges. Advanced preprocessing techniques are employed to eliminate false positives and handle long texts during multi-label classification.

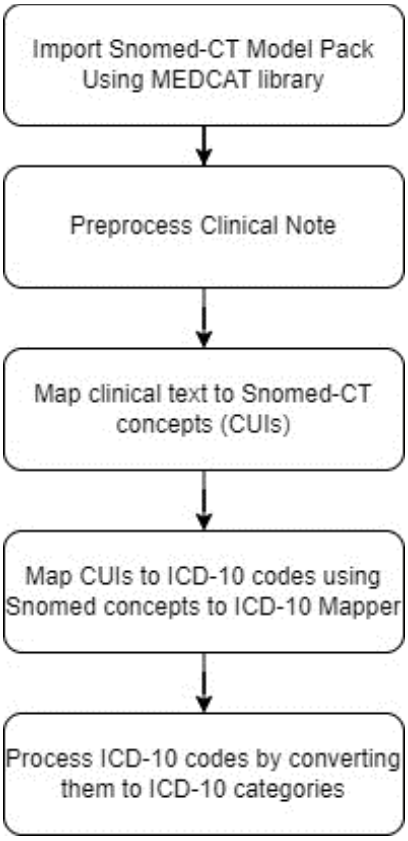
3.3 Process

3.3.1: Preprocessing

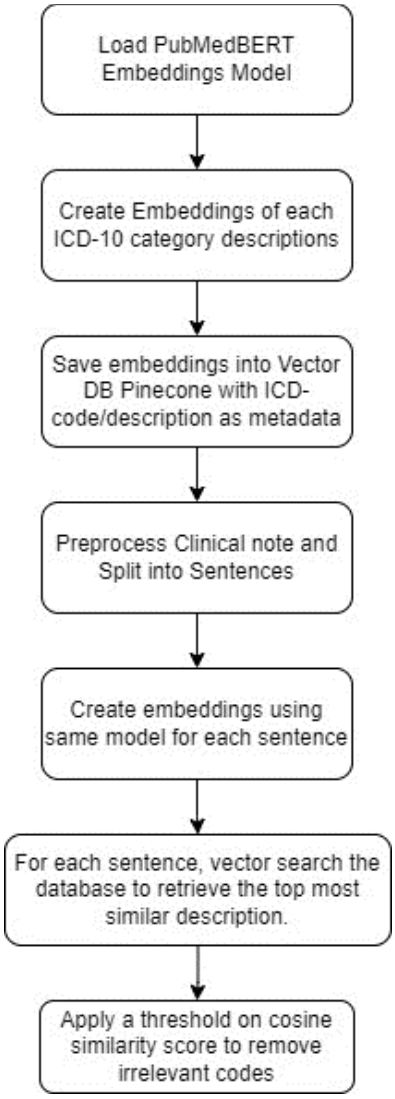
The preprocessing of this project was a crucial step because free text can vary widely, and there are fixed preprocessing techniques for all clinical text types, with additional techniques focused on clinical text from the MIMIC IV dataset. Let's review a few steps taken to ensure correct preprocessing:



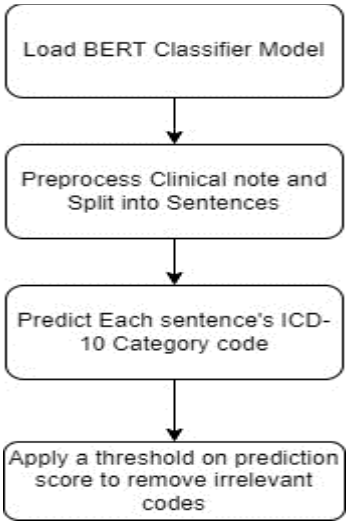
3.3.2: Using MEDCAT and SNOMED-CT



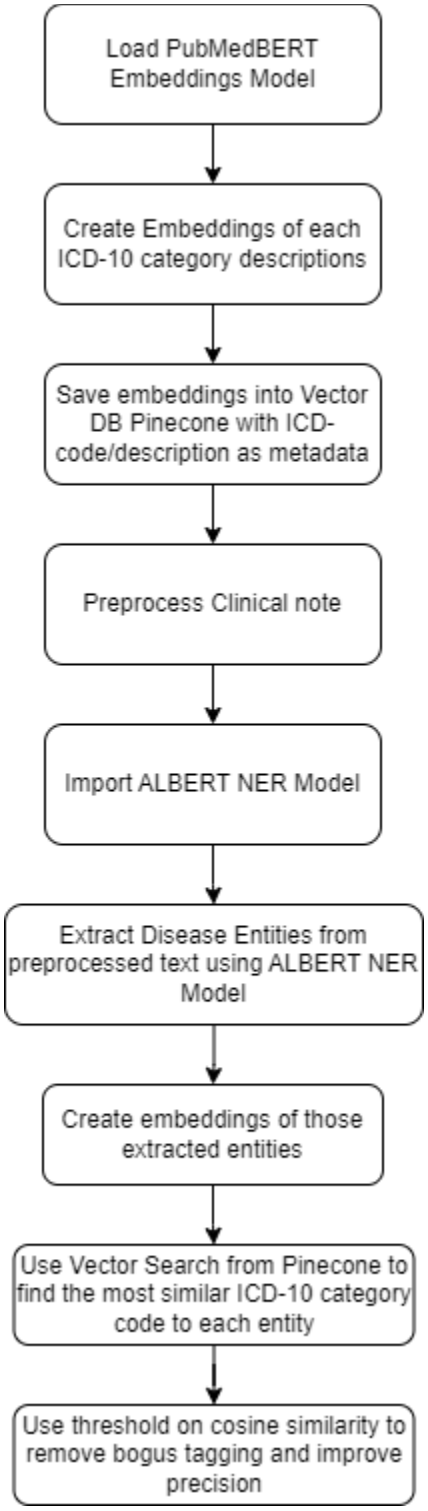
3.3.3: Using Embedding Model (PubMedBERT)



3.3.4: Using Pretrained BERT Classifier (Emran's)



3.3.5: Using AlBERT NER tagger with PubMedBERT embedding model.



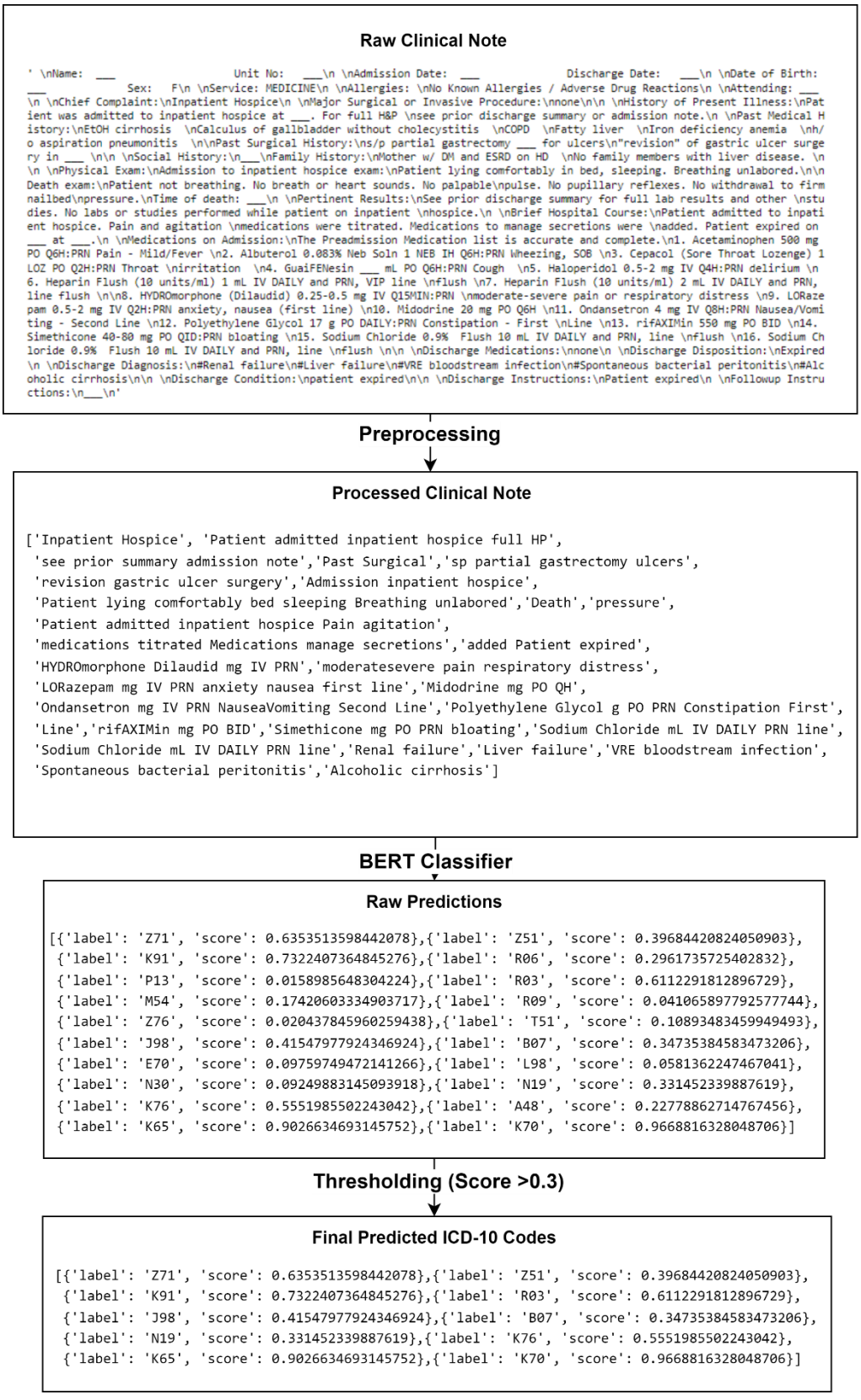
Chapter 4: Experimentation and Results

4.1 Dataset and Preprocessing

The dataset utilized in the series of experiments is MIMIC IV, chosen due to its inclusion of labeled ICD-10 codes. For this set of experiments, the 1000 shortest notes were selected, and all experiments were conducted on this subset. This selection was made because the sample hospital notes from Pakistani hospital were considerably shorter than those in the MIMIC IV dataset, where the average note length was 11,545. In contrast, local hospital's notes averaged a length of 410. The 1000 notes selected for experimentation had an average length of 2739.

Following preprocessing steps outlined in section 3.3.1, the selected notes were subjected to the application of these steps. This resulted in the creation of a list of sentences, which were then input into the classification techniques discussed in other sections for further analysis.

4.2 Example Walkthrough using BERT Classifier



4.3 Experimentations Performed

The experiments were assessed using a sample of 1000 selected clinical notes from MIMIC IV. The chosen evaluation metrics primarily included various forms of Macro Precision, Recall, and F1-scores. The non-zero average metrics represent the Macro-average of non-zero precisions for ICD-10 code categories. The top-50 metrics illustrate the performance for the top 50 most frequently occurring ICD-10 code categories within the 1000 clinical notes. Now, let us proceed to the tables:

4.3.1: MedCAT using SNOMED-CT Model-Pack

Preprocessing	Macro Precision	Macro Recall	Macro F1	Macro Precision Top 50	Macro Recall Top 50	Macro F1 Top 50
Old	**0.266**	**0.261**	**0.218**	**0.399**	**0.193**	**0.230**
New	0.266	0.207	0.197	0.429	0.125	0.170

Preprocessing	Non-Zero Macro Precision	Non-Zero Macro Recall	Non-Zero Macro F1	Non-Zero Macro Precision Top 50	Non-Zero Macro Recall Top 50	Non-Zero Macro F1 Top 50
Old	**0.517**	**0.530**	**0.441**	**0.604**	**0.292**	**0.349**
New	0.608	0.474	0.451	0.671	0.195	0.266

4.3.2: Pretrained BERT Classifier (Emran's Model)

Threshold	Macro Precision	Macro Recall	Macro F1	Macro Precision Top 50	Macro Recall Top 50	Macro F1 Top 50
0.1	0.170	0.177	0.130	0.398	0.112	0.134
0.15	0.179	0.169	0.131	0.413	0.103	0.127
0.2	**0.185**	**0.163**	**0.134**	**0.394**	**0.098**	**0.123**
0.25	0.182	0.156	0.132	0.365	0.090	0.114

Threshold	Non-Zero Macro Precision	Non-Zero Macro Recall	Non-Zero Macro F1	Non-Zero Macro Precision Top 50	Non-Zero Macro Recall Top 50	Non-Zero Macro F1 Top 50
0.1	0.424	0.440	0.323	0.511	0.144	0.171
0.15	0.460	0.434	0.337	0.558	0.140	0.171
0.2	**0.485**	**0.428**	**0.352**	**0.547**	**0.136**	**0.170**
0.25	0.500	0.429	0.363	0.553	0.137	0.172

4.3.3: PubMedBERT Embeddings

Threshold	Macro Precision	Macro Recall	Macro F1	Macro Precision Top 50	Macro Recall Top 50	Macro F1 Top 50
0.3	0.126	0.200	0.109	0.353	0.120	0.135
0.35	0.128	0.199	0.110	0.358	0.119	0.136
0.4	0.131	0.192	0.111	0.367	0.112	0.133
0.45	0.143	0.178	0.117	0.366	0.100	0.126
0.5	**0.151**	**0.161**	**0.117**	**0.365**	**0.078**	**0.104**
0.55	0.153	0.136	0.109	0.315	0.063	0.088
0.6	0.156	0.110	0.100	0.335	0.051	0.077
0.65	0.131	0.078	0.079	0.238	0.034	0.053
0.7	0.116	0.060	0.065	0.208	0.023	0.038

Threshold	Non-Zero Macro Precision	Non-Zero Macro Recall	Non-Zero Macro F1	Non-Zero Macro Precision Top 50	Non-Zero Macro Recall Top 50	Non-Zero Macro F1 Top 50
0.3	0.296	0.469	0.256	0.452	0.153	0.173
0.35	0.301	0.468	0.259	0.459	0.152	0.174
0.4	0.316	0.461	0.266	0.470	0.144	0.170
0.45	0.363	0.452	0.297	0.495	0.136	0.171
0.5	**0.412**	**0.440**	**0.318**	**0.521**	**0.112**	**0.149**
0.55	0.489	0.435	0.350	0.584	0.116	0.162
0.6	0.570	0.405	0.365	0.697	0.107	0.161
0.65	0.635	0.379	0.380	0.700	0.101	0.157
0.7	0.672	0.345	0.374	0.742	0.082	0.137

4.3.4: ALBERT NER with PubMedBERT Embeddings

Threshold	Macro Precision	Macro Recall	Macro F1	Macro Precision Top 50	Macro Recall Top 50	Macro F1 Top 50
0.3	0.115	0.146	0.095	0.261	0.055	0.086
0.35	0.116	0.146	0.093	0.249	0.066	0.083
0.4	0.120	0.143	0.093	0.280	0.065	0.084
0.45	0.123	0.137	0.094	0.294	0.062	0.079
0.5	**0.133**	**0.126**	**0.095**	**0.307**	**0.055**	**0.073**
0.55	0.136	0.117	0.095	0.221	0.046	0.061
0.6	0.130	0.103	0.089	0.259	0.040	0.055

Threshold	Non-Zero Macro Precision	Non-Zero Macro Recall	Non-Zero Macro F1	Non-Zero Macro Precision Top 50	Non-Zero Macro Recall Top 50	Non-Zero Macro F1 Top 50
0.3	0.336	0.419	0.270	0.407	0.098	0.134
0.35	0.339	0.426	0.273	0.402	0.107	0.134
0.4	0.354	0.424	0.276	0.451	0.105	0.135
0.45	0.383	0.425	0.293	0.490	0.103	0.132
0.5	**0.440**	**0.419**	**0.315**	**0.548**	**0.098**	**0.130**
0.55	0.486	0.418	0.341	0.526	0.111	0.144
0.6	0.525	0.419	0.360	0.648	0.101	0.137

Chapter 5: Discussion

This project's primary goal was to do a comparative analysis of assigning multiple ICD-10 codes to an input clinical note using pretrained language models. Following several experiments, it was decided to map it to ICD-10 code categories, reducing the scope from 70,000 codes to 24,000 categories (first three digits). The research for the project centered on comparing the performance of existing libraries, such as MEDCAT's SNOMED-CT, to transformer-based approaches such as BERT, GPT, and others.

Initially, GPT was used in conjunction with Pinecone vector search for few-shot learning. Several prompts were tried, but the results were unimpressive. As a result, the GPT methodology was abandoned, with the primary focus shifting to comparing MEDCAT's SNOMED-CT with pre-trained LLMs such as different BERTs and NER-taggers. The testing set included the first 1,000 notes from the MIMIC IV dataset, which had already been labelled with ICD-10 codes. Macro-F1, Precision, Recall, and a derived metric that excluded ICD codes with zero correct predictions were used as comparison metrics.

Following the selection of the testing set and metrics, the SNOMED-CT model-pack from MEDCAT was imported and tested on the preprocessed clinical notes. It outperformed RoBERTa-PM with a Macro-F1 score of 0.218 and an F1@50 score of 0.230. Emran's BERT model [2] was then used, yielding a Macro-F1 score of 0.134 and an F1@50 score of 0.123. Despite being inferior to MEDCAT, it outperformed RoBERTa-PM [4].

Additionally, another methodology was tried, applying Emran's BERT model [2] to analyze the 1,000 clinical notes, using these indicators as a standard. After a thorough preprocessing phase, the notes were segmented into sentences and the model was run. To exclude predictions that were unclear, thresholding was used to the prediction scores. Consequently, the F1@50 score was 0.123 and the Macro-F1 score was 0.134. The model clearly performed worse than MEDCAT, but it was still better than the metrics in the PLM study, which gave a Macro-F1 of 0.104. It is important to emphasize that Emran's model was not trained on long clinical notes; the shortest 1,000 notes in MIMIC IV averaged 2,300 words, while the notes in that study were less than 50 words. Emran's paper's remarkably high F1 metrics are explained by how concise the notes used were, precisely matching the official ICD-10 descriptors. Several preprocessing methods, such as systematically breaking down notes into sentences for the model's sequential processing, were used to fully use Emran's trained model.

After then, a different approach utilizing PubMedBERT embeddings was investigated. After breaking up the text into sentences and applying the same preprocessing method, sentence-level embeddings were compared to ICD-10 category descriptions. To get rid of inaccurate predictions, the procedure involved thresholding on cosine similarity. A threshold of 0.5 was used to get the best possible balance between recall and precision. With a Macro-F1 score of 0.151, this strategy outperformed the BERT methodology by a little margin. Its F1@50 score, however, was 0.104—lower than what the BERT technique was able to produce.

Finally, a different approach was undertaken using an NER-based model. Multiple NER models were tested, and the only one that yielded reasonable results was ALBERT-Medical-NER from Hugging Face, for which the paper is unavailable. In this method, disease entities were extracted from processed clinical notes, and their embeddings were generated using the PubMedBERT model. The embeddings at the entity level were then

compared with ICD-10 category descriptions, and thresholding on cosine similarity with a threshold of 0.5 was employed to achieve the optimal balance between precision and recall. However, this methodology resulted in a Macro-F1 of 0.095 and an F1@50 of 0.073, marking it as the least effective among all the approaches tested in this project. The primary issue lies in the small lengths of the extracted disease entities compared to the description embeddings. For instance, a comparison between the embedding of "stroke" and "Stroke, not specified as hemorrhage or infarction" would inevitably yield a low score.

Overall, it is evident that the Emran's BERT model, coupled with advanced preprocessing methodology, stands out as the most effective transformer-based model comparable to MEDCAT's benchmark. Both the benchmark and this methodology surpass the best metrics reported in the PLM paper. It is crucial to acknowledge that these methodologies were exclusively tested on the shortest clinical notes from MIMIC IV, whereas the PLM paper utilized a substantial testing dataset from MIMIC II. This discrepancy could potentially explain the inflation in our metrics compared to that paper. Additionally, a direct comparison with Emran's paper is challenging as their results were based on noticeably short clinical summaries, and certain papers have prioritized micro precision over Macro.

Furthermore, it is noteworthy that most research studies exhibit a Macro-F1 score around 0.1, highlighting the comparability of our methodologies to the latest research trends. The primary factor contributing to the relatively low Macro-F1 scores in both this project and other research is the inherent challenge posed by the multi-label classification problem, involving nearly 24,000 ICD-10 code categories. Taking a Macro-average over the F1 scores inherently results in a smaller value. Nonetheless, achieving comparable performance in such a complex problem is a commendable accomplishment.

Chapter 6: Conclusion

Using 1,000 of the shortest clinical notes from the MIMIC IV dataset, this research rigorously tested many approaches with the goal of streamlining the automation of clinical note mapping to ICD-10 codes. With an amazing Macro F1-score of 0.218, the SNOMED-CT model from the MEDCAT library proved to be the best performer and set a milestone for the project.

Furthermore, with a Macro-F1 score of 0.134, Emran's BERT model combined with sophisticated preprocessing performed better than any of the transformer-based approaches investigated. It is important to stress that this low score is significant given the multi-label classification problem with about 24,000 classes. Such a performance is a remarkable achievement because of the complexity of the problem.

This project's potential impact in a clinical setting is highlighted by its real-world implementation, which also represents a major advancement in improving the accuracy and efficiency of managing massive volumes of clinical data.

References

[1] Nguyen, Anthony N., Donna L. Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael Lawley and Damian J. Green. “Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings.” AMIA ... Annual Symposium proceedings. AMIA Symposium 2018 (2018): 807-816.

[2] Al-Bashabsheh, Emran, Ahmad Alaiad, Mahmoud Al-Ayyoub, Othman Beni-Yonis, Raed Abu Zitar, and Laith Abualigah. "Improving clinical documentation: automatic inference of ICD-10 codes from patient notes using BERT model." The Journal of Supercomputing (2023): 1-25.

[3] Amin, Saadullah, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. "MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT." In CLEF (Working Notes), pp. 1-15. 2019.

[4] Huang, Chao-Wei, Shang-Chi Tsai, and Yun-Nung Chen. "PLM-ICD: automatic ICD coding with pretrained language models." arXiv preprint arXiv:2207.05289 (2022).

[5] “International Classification of Diseases.” Encyclopædia Britannica. Accessed December 29, 2023. https://www.britannica.com/topic/International-Classification-of-Diseases.

[6] Ahmed, Joinal. “Automatic ICD-10 Code Assignment to Consultations Using Deep Learning.” Halodoc Blog, June 29, 2023. https://blogs.halodoc.io/automatic-icd-10-code-assignment-to-consultations-using-deep-learning/.

[7] Kraljevic, Zeljko, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio et al. "Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit." Artificial intelligence in medicine 117 (2021): 102083.

[8] Zeljko. “MedCAT: Introduction - Analyzing Electronic Health Records.” Medium, April 23, 2020. https://medium.com/@w_is_h/medcat-introduction-analyzing-electronic-health-records-e1c420afa13a.

[9] Johnson, Alistair, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. "Mimic-iv." PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/ (accessed August 23, 2021) (2020)

[10] Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing." ACM Transactions on Computing for Healthcare (HEALTH) 3, no. 1 (2021): 1-23.