NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting Cardiovascular Disease from Unstructured Clinical Notes

Application of Advanced Natural Language Processing on MIMIC-IV database.

Kauser Al Rashid

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Predicting Cardiovascular Disease from Unstructured Clinical Notes

Application of Advanced Natural Language Processing on MIMIC-IV database.

by

Kauser Al Rashid

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Carina Isabel Andrade Albuquerque (PhD), NOVA IMS. Universidade Nova de Lisboa

May, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Kauser Al Rashid, May 30, 2024]

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of death globally, significantly impacting mortality and morbidity individual across different demographics. The aim of this study is to leverage attention-based Natural Language Process (NLP) models to predict severe forms of CVD from unstructured clinical notes using discharge summaries of patients in MIMIC-IV dataset. Through a comparative analysis of various models that included LSTM, BERT, clinicalBERT and Clinical LongFormer, as well as modified versions of BERT and clinicalBERT, this research finds that attention-based models outperform traditional deep learning models in handling long and complex unstructured clinical notes, and therefore make better predictions. The best performing model identified in this study is BERT (sliding window), as this model was most accurate (Accuracy: 0.73), well-balanced in predictions (F1-Micro: 0.80) and excelled at correctly predicting specific CVD (AUC: 0.83). Although there are some limitations, this study demonstrates the predictive power of advanced attention-based models in healthcare, which would enable better disease predictions and timely interventions to reduce mortality and morbidity due to CVD.

KEYWORDS

Electronic Health Records (EMRs); Clinical Notes; Natural Language Processing; Transformer-based Methods; Cardiovascular Diseases.

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Natural Language Processing
ATTR-CM	Transthyretin Amyloid Cardiomyopathy
AUC	Area Under the Curve
AUPRC	Area Under the Precision Recall curve
AUROC	Area Under the Receiver-operating Characteristics curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional LSTM
BOWs	Bag of Words
CAD	Coronary Artery Disease
CDC	Center for Disease Control and Prevention
CNN	Convolutional Neural Network
CRT	Cardiac Resynchronization Therapy
CVD	Cardiovascular Disease
DL	Deep Learning
ED	Emergency Department
EHRs	Electronic Healthcare Records
GP	General Practitioner
ICD	International Classification of Diseases
ICU	Intensive Care Unit
LSTM	Long short-term Memory
MI	Myocardial Infraction
ML	Machine Learning
NLP	Natural Language Processing
NVAF	Nonvalvular Atrial Fibrillation

PAD	Peripheral Artery Disease
RAG	Retrieval Augmented Generation
RFAB	Risk factor Attention-based Model
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pre-training Approach.
SVM	Support Vector Machine
TF-IDF	Term Frequency and Inverse Document Frequency.
TIA	Transient Ischemic Attack
WHO	World Health Organization

1. INTRODUCTION

1.1. BACKGROUND AND RATIONALE

Cardiovascular disease (CVD) is the leading causes of death across the globe. Data published by Center for Disease Control and Prevention (CDC) indicates that 1 in 5 deaths in the USA in 2021 was due to CVD (CDC, 2023). Moreover, CVD remains the leading cause of death in the USA across gender, sex, ethnic and racial groups (CDC, 2023). Similarly, 1 in 4 deaths in the UK are caused by CVD (Raleigh et al., 2022). Furthermore, CVD is also the leading cause of disability, morbidity, and health inequalities in the UK (Raleigh et al., 2022). The pattern also holds for Portugal, as data published by the Instituto Nacional de Estatistica (2023) indicates that CVD is responsible for 25.6% of all deaths in Portugal in 2021. Furthermore, the World Health Organization (WHO) predicts the number of death due to CVD will likely to increase due to increase in obesity and various unhealthy lifestyles are becoming more prevalent across the globe, especially in the developed countries (WHO, 2021). There are a number of risk factors associated with CVD that include Smoking, Stress, Diabetes, Obesity, family history of CVD and various medications (Cosselman et al., 2015; Koene et al., 2016; Joseph et al., 2017; Yusuf et al., 2020). Detecting the risk factors for CVD manually is time-consuming, resource-intensive, and error-prone, owing to a number of reasons including high dimensionality of data and majority of the risk factors are shrouded in the clinical notes (Abrahao et al., 2017; Houssein et al., 2023). Therefore, as argued by Chokwijitkul et al. (2018), application of Machine Learning (ML) models would enable clinicians to identify these risk factors and predict severe CVD in a timely manner that can guarantee effective interventions.

Electronic Health Records (EHRs), which are digital compilations of patient medical history and clinical assessments, are becoming increasingly prevalent due to their ability to improve quality of care for the patients and to lower costs for healthcare providers (Abrahao et al., 2017). However, beyond improving quality of care and lowering healthcare costs, EHRs have become essential for clinical research (Cowie et al., 2017; Houssein et al., 2023). EHRs hold structured data that include laboratory results, diagnosis and prescriptions, as well as large volume of unstructured clinical notes in narrative text formats (Rikard et al., 2020; Sedlakova et al., 2023). These notes offer a complete picture of a patient's condition (Sedlakova et al., 2023), however, the narrative form of the notes pose considerable challenges for clinical research (Houssein et al., 2023).

Natural Language Processing (NLP) helps to address some of those challenges by converting unstructured clinical notes into structured format that can be used for clinical research (Abrahao et al., 2017; Sedlakova et al., 2023). A number of studies (Abraoho et al., 2017; Husseine et al., 2023; Rikard et al., 2022; Zhang and Cao, 2023) have demonstrated effectiveness of such approach. This research explores some of the key studies with emphasis on studies exploring the efficacy of attention-based NLPs in clinical notes research.

1.2. RESEARCH AIM AND OBJECTIVES

The aim of this thesis is to apply attention-based NLP models on the unstructured clinical notes to predict severe form of CVD. The thesis therefore attempts to achieve the following objectives:

- **RO1:** To introduce and discuss various relevant NLP models on unstructured notes to identify CVD and other similar diseases from the existing literature.
- **RO2:** To conduct comparative analysis and present the performances of various NLP models in predicting CVD from unstructured notes.
- **RO3:** To develop/propose an NLP model that is effective and efficient in detecting severe forms of CVD.

1.3. RESEARCH CONTRIBUTION

The initial survey of the literature indicates several gaps and limitations in the existing literature. Such gaps and limitations include lack of effective attention-based models employed to predict disease risk from unstructured data despite vast amounts of such clinical data being available. Furthermore, limited research indicates lack of temporal dynamics integration and absence of advanced attention mechanisms. This study attempts to fill these gaps in the literature by exploring advanced attention-based models while also exploring ways to integrate temporal dynamics (progression of timeline and age) into the models. The model proposed in this study would help to predict severe CVD among patients, which would enable healthcare professionals to make timely interventions.

1.4. THESIS STRUCTURE

This research dissertation contains 6 chapters. Following this introductory chapter (Chapter 1) that outlines the background and rationale of the research, as well as the objectives and contributions of this research, the remaining 5 chapters are as below.

- **Chapter Two** presents the definition of key concepts and related works on those concepts from the survey of the existing literature.
- **Chapter Three** presents the methodology used for this research dissertation. Detailed steps and processes related to data gathering, data preprocessing and implementation of NLP models on unstructured data.
- **Chapter Four** presents the findings of the empirical study undertaken for this research as well as the proposed NLP model.

- **Chapter Five** presents the analysis and discussion of the results that were outlined in chapter four. Moreover, the performance of the proposed model is also discussed in this chapter.
- **Chapter Six** concludes the thesis, outlines the limitations of this thesis and provides recommendations for future studies.

2. LITERATURE REVIEW

2.1. CHAPTER INTRODUCTION

This chapter presents the findings from the survey of the literature, and it is divided into two main sections. The first section presents the discussion of Natural Language Processing with an emphasis on the current state of NLP on Electronic Healthcare Records (EHRs). The second section is focused on cardiovascular disease (CVD), and various indicators and risk factors associated with CVD that are fundamental in predicting severe forms of CVDs in patients.

2.2. NATURAL LANGUAGE PROCESSING

NLP is a subfield of Artificial Intelligence (AI), which can be defined as "a collection of computational techniques for automatic analysis and representation of human languages" (Chowdhary, 2020, p. 604). The concept of NLP emerged in 1950s as an intersection of AI and Linguistics with the aim of enabling computers to understand and generate human language (Nadkarni et al., 2011). NLP can be divided into two main subfields: Natural Language Understanding (NLU) and Natural Language Generation (NLG) (Chowdhary, 2020; Khurana et al., 2022). NLU is concerned with comprehension of human language (written or spoken), while NLG generates human-like text from structured data. Given the purview of this thesis, only NLU is discussed in the subsequent sections.

As discussed by Cervantes et al. (2020), early days of NLP mainly involved statistical techniques such as Support Vector Machine (SVM). However, during the early 2000s, Deep learning (DL) began to take prominence due to significant advancement of computer hardware (Khurana et al., 2022). DL is a subset of ML, the development of which was inspired by the workings of human brain (Cervantes et al., 2020). Although DL became popular in 2000s, the origin goes back to 1943 when McCulloch and Pitts (1943) presented their mathematical model that used binary neurons connected in such a way that formed circuits, which were capable of performing logical operations. Their model mimics how biological neurons work inside the brain, where one neuron receives information, which it then processes and sends to the next neuron (Piccinini, 2004). Singular neurons themselves do not have any learning mechanisms of their own, but collectively they form the foundation for the algorithms (Khurana et al., 2022).

Perceptron was the next breakdown in NLP (Piccinini, 2004). Proposed by Rosenblatt in 1958, this is a more advanced version of the neuron proposed by McCulloch and Pitts (Lefkowitz, 2019). The Perceptron has learning capabilities (Lefkowitz, 2019). Over the following decade, Ivakhneno (1963) and his team lead the development of a set of algorithms, known as "Group Method of Data Handling", which are used to solve a number of advanced real-world

problems. These algorithms have since been recognized as the first-ever Artificial Neural Networks (ANNs) (Khurana et al., 2022).

During the 1980s, Fukushima proposed Neocognitron, a neural network for recognizing visual patterns (Nadkarni et al., 2011). Neocognitron has been widely recognized as the inspiration of the Convolutional Neural Networks (CNNs) (Nadkarni et al., 2011). Fukushima (2003) claims that Neocognitron was able to differentiate and classify handwritten numerals successfully, which an accuracy of 98.6%.

Building on the work of Fukushima, Rumelhart et al. in 1968 made significant contributions to DL. Their paper introduced 'Back-Propagation' (Wythoff, 1993), a method that is fundamental for training complex models, and in the process, they laid the groundwork for Recurrent Neural Networks (RNNs), the foundations for many of today's RNNs including Long Short-Term Memory (LSTM). During the 1970s and 1980s, DL methods were refined and applied in diverse fields including the healthcare industry. However, as Nadkarni et al. (2011) pointed out the adaptation was very slow and unenthusiastic given architectural limitations of these models, which were further handicapped by the computer hardware.

The early part of the 21st century enabled DL models to become more effective due to algorithmic advancements, increased computational power and explosion of digital data in the form of images, texts, audio and video (Gorriz et al., 2023). This enabled various industries including healthcare industry to adapt DL models for various tasks including medical image analysis. The application of various ML models during 2010s has become more prevalent (Gorriz et al., 2023). Examples including application of Support Vector Machine (SVMs) in diagnosing cancer based on gene expression, identification or various disease risk factors and Classification of Medical images, and the application of Decision Trees to make diagnostic decisions based on symptoms (Chervonenkis, 2013). Examination of these use cases indicates the confinement of DL and ML models on structured data. This is because the DLs and MLs during this decade despite their improvements were unable to handle the Data Complexity that is associated with understanding abbreviations, jargons and complex grammar that are inherent in EHR data (Holmes et al., 2022). Furthermore, Holmes et al. (2022) added, limited availability of data, black-box nature of DL models and lack of computational power have limited the adaptation of DLs and MLs during 2010s.

The complexity of understanding human language, especially consideration of context remained a difficult problem for DLs during 2010s despite some of the improvements made with the invention of LSTM (Gorriz et al., 2023). However, one of the major breakthroughs in NLP came in the latter part of 2010s following the publication of a landmark paper by Vaswani et al. (2017), titled, "Attention is all you Need". The transformer technique proposed by Vaswani et al. (2017) is able to deal with sequential data and long-term dependency. The intuition behind this architecture was to address the problem of handling long text sequences by removing recurrent operations as previous neural networks and replacing them with self-attention.

Self-attention is based on the attention mechanism proposed by Bahdanau et al. (2016), mainly for improving the performance of machine translation networks. The significance of self-attention mechanism is that it pays attention to other words in a given sequence. The attention mechanism measures the attention between predicted words with respect to their input words, while self-attention measures the attention between all input words in respect to each other. The transformer architecture (shown in figure 2.1) is composed of an encoder block and a decoder block. Each block contains attention, residual connections as well as point-wise-feed-forward network.

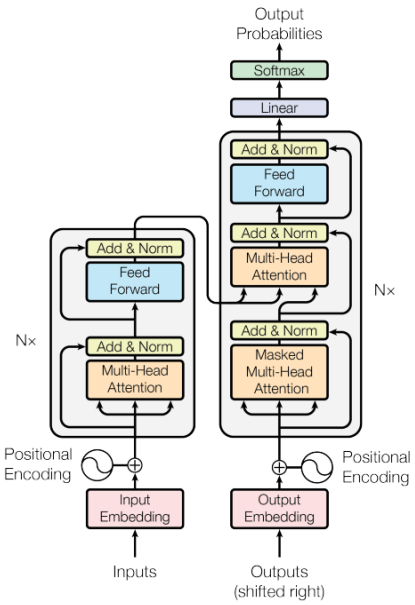

Figure 2.1 – The Transformer – Model Architecture (Viswani et al., 2017)

In RNNs, positional information is inherently captured as the input is processed sequentially over time. However, when the recurrence mechanism is removed, this positional information is lost. To address this issue, the Transformer model incorporates a positional encoder. The positional encoder assigns positional information to each word in the sequence based on its position within the input embedding. This enables the model to maintain awareness of the position of each word in the sequence.

$$PE(pos, 2i) = \sin\left(\frac{pos}{1000^{\left(\frac{2i}{d_{model}}\right)}}\right)$$

$$PE(pos, 2i) = \cos\left(\frac{pos}{1000^{\left(\frac{2i}{d_{model}}\right)}}\right)$$

Equation 2.1 Mathematical Expressions for encoding for inputs.

Equation 2.1 demonstrates the mathematical expressions used for encoding the inputs. Different words are distinguished by their positional index, ensuring unique positional identification of each word. Each dimension "i" is given a specific frequency, which then enables the positional encoder to capture both words that are nearby and words that are further apart.

Multi-Head Attention consists of several self-attention operating in parallel (as shown in figure 2.1). The Transformer model divides the computation across different heads, processes them simultaneously, and then concatenates the results into a single attention head. Equation 2 presents the formula for scaled dot-product attention and illustrates the multi-head attention mechanism, as described by Vaswani et al. (2017). In this process, Q stands for the query, K for the key, and V for the value. This is analogous to querying a database: a user sends a query, the database matches it against the keys, and returns the corresponding values. In self-attention, with the input as the focus, Q, K, and V all receive the same input.

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$MultiHead\ (Q, K, V) = Concat(head_1, \dots, head_h)W^o$$

$$\textit{Where, } head_i = Attention\ (QW^Q{}_i, KW^K{}_i, VW^V{}_i)$$

Equation 2.2 Attention and Multiheaded Attention in Transformer Architecture

The output of the multi-head attention is combined with the input via a residual connection and then normalized before being passed into a feed-forward network. The decoder operates similarly to the encoder but focuses on computing a representation of the output embedding. It utilizes Masked Multi-Head Attention instead of standard multi-head attention. In Masked Multi-Head Attention, attention is computed only between words that precede the current predicted word in the sequence, ensuring that future words are not taken into account.

There are several models based on transformer architecture. Given the relevancy, this study mainly focuses on Bidirectional Encoder Representations from Transformers (BERT). Developed by Delvin et al. (2018), this model however only makes use of the encoder layer. The encoder layer in BERT consists of multiple identical layers (usually 12 or 24). Each layer comprises multi-head attention followed by a feed-forward network, with residual connections and layer normalization applied at each sub-layer.

The final output of BERT is a sequence of contextualized embeddings for each token, which can be used for various NLP tasks, such as classification, named entity recognition, and question answering. For multilabel classification, BERT uses a fully connected (dense) layer with a sigmoid activation function as it allows for the prediction of multiple independent probabilities, which is suitable for multilabel classification. Throughout this dissertation, the

terms ‘transformer architecture-based models’ and ‘attention-based models’ are used interchangeably.

2.2.1. Current State of NLP in EHRs

NLP is being used widely across different disciplines, and performance of NLP varies depending on the field and complexity of the tasks. This section presents how NLP models are being used in healthcare, especially in analyzing unstructured notes in EHRs, and the performance of those models from established literature.

In order to locate the related work, the metadata of articles in three leading scholarly databases in science, medicine, and computer science: PubMed, ACM Digital Library and IEEE Explore were searched. The search terms used for each database are shown in Appendix A. Furthermore, publications between 2017 and 2023 were selected to ensure that the literature review captures the rapid advances in NLP as recommended by Turchioe et al. (2022). Moreover, as recommended by Boren and Moxley (2015), literature review involving clinical practices and interventions should include recent literature to capture new and improved clinical practices, interventions, and phenotype detections.

Besides the selected timeframe, studies are also excluded if they (a) are not published (or available) in English, (b) are not available in full, (c) are duplicates, (d) are not reviews (such as Systematic Literature Review), and (e) are not focused on application of NLP tools. The application of these 5 criteria for exclusion yielded 213 articles from 727 articles that were identified in the initial search (Appendix A). The abstracts of these 213 were then screened to determine their relevancy to the research scope and research questions (n = 78). Finally, the full text of the articles was reviewed to exclude articles without cardiology focus or articles that failed to detail NLP tools/methods employed by them (n = 19). Finally, these 19 articles were appraised using eight Indicators (shown in Table 2.1 of Quality proposed by Koleck et al. (2019). This research only includes articles that satisfy at least 5 out of 8 indicators shown in table 2.1. The use of such benchmarks (satisfying 5 out of 8 indicators) is arbitrary as quality criteria for NLP research is still evolving (Turchioe et al., 2022). The PRISMA flow diagram (figure 2.2) illustrates the process of study selection for this literature review.

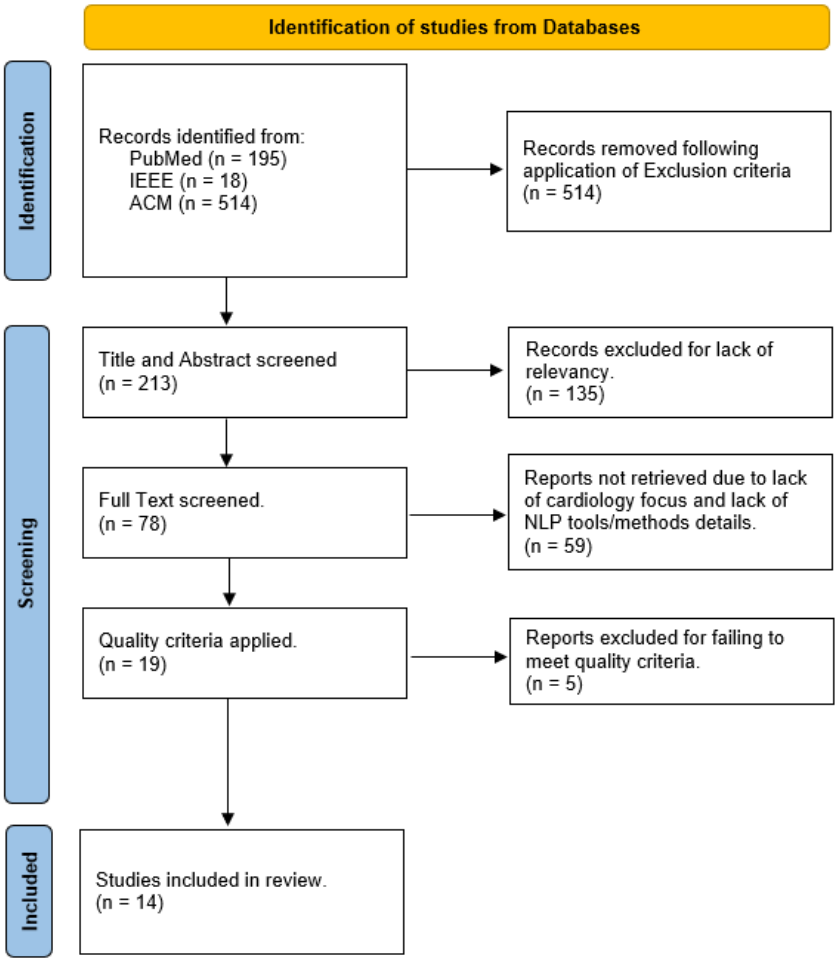

Figure 2.2 PRISMA flow diagram.

Table 2.1 Quality Indicators for NLP research studies

Indicators
1. Clearly Defined purpose
2. Symptoms as primary outcomes
3. Approach Adequately described
4. Number of Documents specified.
5. Number of Patients specified.
6. Patient demographic information reported.
7. Evaluation Metrics reported.
8. Inclusion of Comparative evaluation

The final 14 studies selected for review in this research can be divided into two groups based on their aims – (a) Predicting CVD and/or CVD treatment outcomes, and (b) Identifying and/or Extracting events related to CVD. The table below presents these studies and their aims.

Table 2.2 List of Studies included in this Review and their Aims.

Group	Studies	Aims
Predicting CVD and/or CVD treatment outcomes	Hu et al. (2019)	Predicting Cardiac Resynchronization Therapy (CRT) outcomes.
	Liu et al. (2019)	Predicting Hospital Readmission.
	Qiu et al. (2022)	Predicting CVD from risk factors and EHR texts
	Anetta et al. (2022)	Predicting CVD diagnosis from patients' medical history
	Guazzo et al. (2023)	Predicting hospitalization due to CVD among diabetes patients
Identifying and/or Extracting events related to CVD	Adekkannattu et al. (2019)	Extracting Cardiac concepts
	Viani et al. (2019)	Extracting key clinical events
	Zhan et al. (2021)	Extracting Diagnostic codes for CVD
	Singh et al. (2022)	Extracting quantitative measures from Magnetic Resonance Imaging (CMR) reports
	Moya et al. (2023)	Detecting Transthyretin Amyloid Cardiomyopathy
	Elkin et al. (2021).	Identifying various CVDs
	Houssein et al. (2023)	Identifying CVD risk factors in clinical texts
	Roberts et al. (2016)	Identifying CVD risk factors in clinical texts
	Chokwijitkul et al. (2018)	Identifying CVD risk factors in clinical texts

The first set of studies (Anetta et al., 2022; Guazzo et al., 2023; Hu et al., 2019; Liu et al., 2019; Qiu et al., 2022) focused on predicting CVD among the patients from the clinical notes or predicting the outcomes of CVD treatments on the patients from textual notes. Hu et al. (2019) in their study investigated the potential of ML to improve patient selection for CRT, which is a therapy used to improve heart functions of patients with medically refractory systolic Heart Failure (HF) and Left Ventricular Dyssynchrony (LVD). The researchers applied a Word2Vec continuous bag-of-words model on unstructured clinical notes from EHR and managed to achieve 79% accuracy of predicting success of CRT, compared to 59% of success using traditional clinical risk factors alone. It is to note that their model incorporated structured data from EHR along with unstructured notes.

Liu et al. (2019) in their study focused on examining the effectiveness of DL to predict heart failure readmission from clinical notes. The authors applied CNN on the discharge summary notes of hospitalized patients due to HF from MIMIC-III database and achieved F1-score of 0.733 for 30-day readmission prediction and F1-score of 0.756 for general readmission prediction. This DL model significantly outperformed the traditional ML model based on Random Forest. Furthermore, using chi-square test the authors also identified the most

important features that CNN model used to predict readmissions. These features include age, sex, race, comorbidities, and medication adherence.

Qiu et al. (2022) in their study proposed a Risk Factor Attention-based Model (RFAB) to predict CVD. The researchers used BiLSTM-CRF to identify CVD risk factors from patient medical records, which are then used as inputs along with texts from EHR to predict CVD for patients. Their model outperformed (F1-score = 0.9586) other models based on CNN (F1-score = 0.9128), LSTM (F1-score = 0.8217) and SVM (F1-score = 0.9091). Their study contrasted other studies with similar aim as they adopted a mode that fused character sequence with CVD risk factors from EHR text.

Guazzo et al. (2023) in their study used Bidirectional LSTM to predict hospitalization due to CVD from outpatient medical reports of diabetic patients. Their model performed well in predicting hospital admission at infinite time window (F1-score = 0.94 [20% uncertainty]; F1-score = 0.80 [0% uncertainty]), as well as predicting hospital admission within the next 24 months (F1-score = 0.92 [20% uncertainty]; F1-score = 0.76 [0% uncertainty]). However, the model underperformed in predicting hospitalization within 6-months (F1-score = 0.64 [20% uncertainty]; F1-score = 0.50 [0% uncertainty - 6]). The authors argued that the underperformance at shorter time frame was due to most patients scheduling clinic visits at > 1-year intervals. Also, the number of false positive increases as prediction time-window gets shorter.

The Polish study conducted by Anette et al. (2022) used data from 50,000 patients in the cardiology department in Medical University of Siesia, Poland. Unlike, the three studies discussed above, their study used LLM based on transformer architecture – Polish RoBERTa. The choice of which was dependent on the textual data being Polish, and authors claimed that previous studies have shown that Polish RoBERTa outperformed other LLM trained on Polish corpora. The results of their study are mixed, as their model performed well in identifying certain CVD such as Acute Myocardial Infraction (ICD10 – I21), their model underperformed in identifying Chronic IHD (ICD10 – I25). The authors blamed the constraints of text modality in medicine for their poor performance of their model in identifying some of the CVD categories.

The second set of studies are focused on extracting concepts or identifying risk factors from text data using NLP tools/methods. The first of these studies was conducted by Adekkannattu et al. (2019) using data from MIMIC-III and 2 Academic Medical Centres in the USA. In their study, they assessed the efficacy of EchoExtractor, an NLP pipeline that utilizes Leo infrastructure to extract clinical measurements from echocardiogram reports. The study found that EchoExtractor was able to extract key measurements from echocardiogram reports across all databases with F1 score > 0.90. Their study highlights the ability of NLP models to extract key cardiology measurements from text data within considerably shorter amount of time compared to manual extraction.

The study of Viani et al. (2019) had similar aim as they employed RNN to extract key clinical events from medical reports of cardiology patients. The authors trained their model using 75 annotated medical reports containing 4,365 relevant events (such as ECG, admission, and Brugada Syndrome). Compared to ruled-based Dictionary Lookup approach (F1 score = 0.637), and conventional ML approach such as SVM (F1 score = 0.813), Viani et al.'s (2019) RNN model, Bi-LSTM-CRF-CHAR (F1 score = 0.886) significantly improved the ability of healthcare providers to extract key clinical events for cardiology patients. However, the best performance was achieved by merging Bi-LSTM-CRF-CHAR with Dictionary Lookup (F1 score = 0.901).

The study of Singh et al. (2022) had similar aim as the researchers in this study attempted to extract quantitative measurements (such as Left ventricular end diastolic volume and Aortic root dimension) from Cardiac Magnetic Resonance Imaging (CMR) reports using LLMs that include PubMedBERT, Bio+DischargeSummaryBERT, SapBERT and BERTLarge. The best-performing model was BERTLarge that was fine-tuned on replaced decimal numerical representation scheme achieving F1-score of 0.957.

Elkin et al. (2021) in their study employed HD-NLP to identify Nonvalvular Atrial Fibrillation (NVAF), which lead to strokes and death. The author found that the use of structured data along with unstructured data (EHR notes, laboratories, and problem lists) improved identification of NVAF among patients compared to just using structured data (F1-score = 0.964 vs F1-score = 0.686). The authors concluded that the use of structured data along with unstructured data could have prevented 176,537 strokes annually in the USA preventing 10,575 deaths while saving the US economy over $13.5 billion USD.

A number of studies in the literature also used various proprietary NLP algorithms/software to analyze clinical notes. The study of Moya et al. (2023) is one such studies that used LynxCare, an NLP algorithm developed by LynxCare Inc. to detect Transthyretin amyloid cardiomyopathy (ATTR-CM), a fatal cardiomyopathy that is often misdiagnosed or diagnosed late preventing patients from obtaining timely treatment. Using unstructured and semi-structured EHR data from OLV Hospital Aalst in Belgium, this study found that ATTR-CM can be detected in timely manner with high precision (mean precision of 0.9732; F1-score = 0.9841).

However, despite the increasing adoption of attention-based NLP models, these models do not always perform better than conventional models for certain tasks. This was affirmed by Zhan et al. (2021) in their study that focused on extracting diagnostic codes for CVDs (ICD-10 codes) from MIMIC-III and Stanford EHR dataset. Using TF-IDF, this study managed to achieve high (0.9499–0.9915) Area Under the Receiver-operating Characteristics curve (AUROC) and high (0.2956–0.8072) Area Under the Precision Recall curve (AUPRC). On the contrary, Rishivardhan et al. (2020) achieved F1-score of 0.033 and highest precision of 0.025 while attempting to extract ICD-10 codes using BERT, RoBERTa, Electra, and XLNet.

The remaining three studies (Chokwijitkul et al., 2018; Houssein et al., 2023 and Roberts et al., 2016) used the same dataset (2014 i2b2 Heart Disease Risk factors challenge dataset) to identify heart disease risk factors in clinical text. Roberts et al. (2016) employed SVM, while Chokwijitkul et al., 2018 assessed five NN-based NLP architectures (CNN, RNN, GRU, LSTM and Bi-LSTM). SVM outperformed (F1-score = 0.9276) all NN-based NLP models (highest F1-score = 0.9081 [Bi-LSTM]). However, Houssein et al. (2023) in their study outperformed (F1-score = 0.9366) all previous attempts using the same dataset by using BERT and CHARACTER-BERT Embedding stacking. Houssein et al. (2023) concluded that the use of contextual embedding may increase the effectiveness of NLP. The summary of the included studies in this review is shown in table 2.2.

Table 2.3 Summary of the Included Studies.

Author (Pub. Year)	Dataset or Data source	Patient Population	Sample Size	Type of Document	NLP Tools/Methods	Performance
Hu et al. (2019)	Partners HealthCare Research Patient Data Registry	Patient underwent Cardiac Resynchronizatio n Therapy (CRT).	990 patients	Clinical notes (inpatient admission notes, outpatient notes, consultation notes, cardiology reports and others)	Word2Vec continuous bag-of-words model	F1-score = 0.77; accuracy = 0.65; Recall = 0.26; Precision = 0.79
Liu et al. (2019)	MIMIC-III	Patient admitted in the hospital	Not mentioned	Discharge Summary Notes	Convolutional Neural Network (CNN)	F1-score = 0.756 [General Readmission Prediction] F1-score = 0.733 [30-day Readmission Prediction]
Qiu et al. (2022)	Data from Internal Medicine department of a hospital in Gansu, China, and Network Intelligence Research Laboratory of Language Technology Research Centre, Habin Institute of Technology.	Hospitalized patients	Not mentioned	Medical records	Risk factor Attention-based Model (RFAB)	F1-score = 0.96; Accuracy = 0.96; Precision = 0.96; Recall = 0.96

Anetta et al. (2022)	Data from EHR system of Department of Cardiology, Medical Univeristy of Siesia, Poland	Hospitalised Cardiology patients	50,000 patients	Admission (incl. reasons) and medical history, Physical examination at admission, discharge summary and results, medication and recommendations at discharge	BERT and RoBERTa	F1-score ranges from 0.53 to 0.90
Adekkannattu et al. (2019)	MIMIC-III, and 3 academic medical centres: Weill Cornel Medicine, Mayo Clinic and Northwestern Medicine.	Patient underwent Echocardiograms	Not mentioned	Echocardiogram Reports	*EchoExtractor*	F1-score > 0.90 for all 4 datasets
Viani et al. (2019)	Reports from Molecular Cardiology Laboratories of the Istituti Clinici Scientifici Maugeri (ICSM)	Cardiology Patients	Not mentioned	Medical Reports (inc. clinical and family history, results of performed tests, possible medication prescription and diagnosis	Recurrent Neural Network (RNN)	F1-score = 0.90; Recall = 0.92; Precision = 88.2%

Guazzo et al. (2023)	EHR from Diabetic Outpatient Clinic of the University Hospital of Padova (Italy)	Patients with Diabetes	16,292 patients	Medical Records pertaining regular outpatient clinic visits	Bidirectional Long Short-term Memory (LSTM)	F1-score = 0.94 [20% uncertainty - Infinite Time Window] F1-score = 0.80 [0% uncertainty - Infinite Time Window] F1-score = 0.92 [20% uncertainty - 24 months Time Window] F1-score = 0.76 [0% uncertainty - 24 months Time Window] F1-score = 0.79 [20% uncertainty - 12 months Time Window] F1-score = 0.63 [0% uncertainty - 12 months Time Window] F1-score = 0.64 [20% uncertainty - 6 months Time Window] F1-score = 0.50 [0% uncertainty - 6 months Time Window]
Singh et al. (2022)	Enterprise Warehouse of Cardiology (EWOC)	Patients made cardiology clinic visits between 2000 and 2019	9,280 patients	CMR reports	PubMedBERT, Bio+DischargeSummary BERT, BERT(large), SapBERT	F1-score = 0.957 [BERT(large)]
Moya et al. (2023)	EHRs from OLV Hospital Aalst, Belgium (2012 - 2020)	Patients attended OLV Hospital	Not mentioned	Structured Tables and Clinical Notes	*LynxCare*	F1-score = 0.98; Recall = 0.99; Precision = 0.97

Zhan et al. (2021)	Stanford EHR dataset and MIMIC-III	Patients attended Stanford Health Care (Stanford EHR) and Hospitalized patients (MIMIC-III)	133,644 patients (Stanford EHR) 41,127 (MIMIC-III)	Outpatient progress notes	TF-IDF and BOWs	AUPRC (0.2956–0.8072) and AUROC (0.9499–0.9915) [TF-IDF on Stanford EHRs] AUPRC (0.2353–0.8084) and AUROC (0.7952–0.9790) [TF-IDF on MIMIC-III]
Elkin et al. (2021)	Data from University of Buffalo faculty practice's EHR.	Patient visited University of Buffalo's outpatient	96,681 patients	EHR notes, laboratories, problem lists and structured EHR data	HD-NLP	The use of unstructured data along with structured data improves models to predict stroke risk by 32.1%
Houssein et al. (2023)	2014 i2b2 Heart Disease Risk factors challenge dataset	Patients attending various hospitals in the USA	Not mentioned	Clinical notes	BERT and CHARACTER-BERT Embedding stacking	F1-score = 0.9366
Roberts et al. (2016)	2014 i2b2 Heart Disease Risk factors challenge dataset	Patients attending various hospitals in the USA	Not mentioned	Clinical notes	Support Vector Machine (SVM)	F1-score = 0.9276
Chokwijitkul et al. (2018)	2014 i2b2 Heart Disease Risk factors challenge dataset	Patients attending various hospitals in the USA	Not mentioned	Clinical notes	CNN, RNN, GRU, LSTM and BLSTM	F1-score = 0.9081 [BLSTM] F1-score = 0.9046 [GRU] F1-score = 0.9010 [LSTM] F1-score = 0.8900 [RNN] F1-score = 0.8793 [CNN]

2.3. CVD – RISK FACTORS

The World Health Organization (WHO) defined CVDs as "a group of disorders of the heart and blood vessels" (WHO, 2021). Lopez et al. (2023) provided a more comprehensive definition of CVD. According to them, CVD refers to "4 entities [disorders]: Coronary Artery Disease (CAD), Cerebrovascular Disease, Peripheral Artery Disease (PAD), and Aortic Atherosclerosis". CAD occurs when the heart muscle does not receive adequate oxygen-rich blood, causing chest pain due to ischemia, which can lead to heart failure or Myocardial Infraction (MI) (Cassar et al., 2009). Cerebrovascular Disease, also known as Transient Ischemic Attack (TIA) or stroke occurs when blood flow to the brain is interrupted or reduced (Lopez et al., 2023). PAD affects the arteries in the limbs, potentially resulting in Claudication (Lopez et al., 2023). Finally, Aortic Atherosclerosis refers to the disorder that involves the buildup of plaque in the aorta, the largest artery in the body, leading to aneurysm in the thoracic and abdominal regions (Cassar et al., 2023).

The pathophysiology of above-mentioned disorders is not explored in this study given their irrelevance to the purview of this thesis, rather Epidemiology of CVD, in particular, the risk factors associated with CVD are briefly discussed in this chapter. The rationale for identifying these risk factors comes from the works of Elkin et al. (2021) and Houssein et al. (2023). In particular, the study of Houssein et al. (2023) highlighted the efficacy of incorporating Risk Factor indicators in the form of contextual embeddings to improve prediction accuracy.

Based on the brief survey of the literature, the following risk factors are identified for cardiovascular diseases.

Table 2.4 Risk factors of CVD.

Category	Risk Factors	Sources
Metabolic	Pre-hypertension	Farrag et al. (2015)
	Hypertension	Ramsay et al. (2014); Roman et al. (2019); Ruan et al. (2018); Wu et al. (2019); Yusuf et al. (2020)
	Diabetes	Yusuf et al. (2020)
	High Waist-to-hip ratio	Ramsay et al. (2014); Yusuf et al. (2020)
	Obesity	Farrag et al. (2015); Roman et al. (2019); Ruan et al. (2018); Weiss et al. (2018); Yusuf et al. (2020)
	Overweight	Wu et al. (2019); Yusuf et al. (2020)

	Low High-density Lipoprotein cholesterol	Roman et al. (2019); Yusuf et al. (2020)
	High low-density lipoprotein cholesterol	Ramsay et al. (2014); Roman et al. (2019); Yusuf et al. (2020)
	High C-reactive protein	Roman et al. (2019)
	Abnormal Lipid Metabolism	Weiss et al. (2018)
Behavioral	Low to moderate Physical activity	Farrag et al. (2015); Roman et al. (2019); Ruan et al. (2018); Yusuf et al. (2020)
	Smoking	Ruan et al. (2018); Weiss et al. (2018); Wu et al. (2019); Yusuf et al. (2020)
	Heavy (alcohol) Drinking	Roman et al. (2019); Ruan et al. (2018); Yusuf et al. (2020)
	Former (alcohol) Drinking	Wu et al. (2019); Yusuf et al. (2020)
	Poor Diet	Roman et al. (2019); Yusuf et al. (2020)
Demographic	Man	Wu et al. (2019)
	Older adult (> 45 years)	Roman et al. (2019); Wu et al. (2019)
Socioeconomic and Psychosocial	Limited Education	Yusuf et al. (2020)
	Lower Income	Ruan et al. (2018)
	Weaker Grip Strength	Yusuf et al. (2020)
Environmental	Air pollution exposure	Yusuf et al. (2020)
Genetic	Family History of CVD	Tian et al. (2017)

2.4. CHAPTER CONCLUSION

This chapter presented a brief history of NLP, development of key NLP architectures, especially the transformer architecture that has revolutionized NLP in the recent years. Furthermore, this chapter also presented a review of established literature on the application of NLP on the unstructured data from EHR. Although transformer architecture has revolutionized other aspects of NLP, the results from studies that explored the application of this architecture on unstructured clinical data to identify disease risk factors and predict diseases have been

mixed. This was evident in the study of Anette et al. (2022) that highlighted the ineffectiveness of transformer-based NLP architecture, especially when faced with medical text modality. However, the study of Houssein et al. (2023) in their study using contextually embedding managed to outperform other similar studies that utilized traditional DL architecture.

The mixed results and continuous improvement of transformer-based architecture warrant exploration of more refined NLP models that can be tailored to medical contexts. Furthermore, literature review also highlighted that while several studies have focused on identifying risk factors of CVD from the notes, there is a lack of studies that focused on predicting CVD from the unstructured clinical notes. Therefore, this research proposes an effective transformer-based model that is effective in predicting CVD diseases for the patients based on their unstructured clinical notes.

3. METHODOLOGY

3.1. CHAPTER INTRODUCTION

The purpose of this chapter is to present the details of the data used for this research. Besides, this chapter also presents the summary statistics of the data and labels. Furthermore, this chapter also outlines pre-processing steps for EHRs adopted for this research.

3.2. DATASET

Medical Information Mart for Intensive Care (MIMIC) is one of the largest and most extensive publicly available EHR dataset (Liao & Voldman, 2023; Strodthoff et al., 2024). Published by MIT's Laboratory for Computational Physiology (LCP) on January 6, 2023, MIMIC-IV (version 2.2) (hereafter "MIMIC-IV") contains de-identified health records of over 315,000 patients that attended Emergency Department (ED) or an Intensive Care Unit (ICU) of Beth Israel Deaconess Medical Center in Boston, USA between 2008 and 2019 (Johnson et al., 2023). The main MIMIC-IV dataset contains 31 files/tables divided into two folders – hosp and icu. hosp folder includes files/tables related to patient admissions to the ED (admissions.csv.gz), their diagnosis (diagnosis_icd.csv.gz), laboratory measurements (such as labevents.csv.gz) as well as administrative information such as medical administration details (emar_details.csv.gz). Supplementary datasets were also made available by LCP that include MIMIC-IV-Note that includes discharge details (discharge.csv.gz) containing discharge summary of a patient following their discharge from an admission.

3.3. DATA PRE-PROCESSING

Mining medical text, in particular unstructured medical texts come with significant challenges that include spelling errors, ambiguity, and the use of abbreviations and acronyms (Joopudi et al., 2018). Moreover, medical abbreviations vary across different specialties. For instance, EAM stands for External Auditory Meatus in the context of Otorhinolaryngology, but standards for Electro-Anatomical mapping in Cardiology (Locati et al., 2023).

Furthermore, as highlighted by Chai (2022), pre-processing of medical texts can inadvertently remove or alter the meaning of information leading to consequential misdiagnosis for the patients. However, some pre-processing is required given the limitations of resources and often models. For instance, Transformer-based models such as BERT have maximum input sequence of 512 tokens (Delvin et al., 2018), indicating need for some pre-processing.

3.3.1. Data Aggregation

The information required for this research is distributed across two different files/tables of MIMIC-IV. Therefore, data from admissions.csv.gz (table: admissions), diagnosis.csv.gz (table: diagnosis) and discharge.csv.gz (table: discharge) are aggregated so that discharge notes can be matched with admission details and diagnosis of the patients. Moreover, demographic information of the patients is retrieved from patients.csv.gz (table: patients). Summary of relevant fields in each of the above four files are illustrated in figure 3.1.

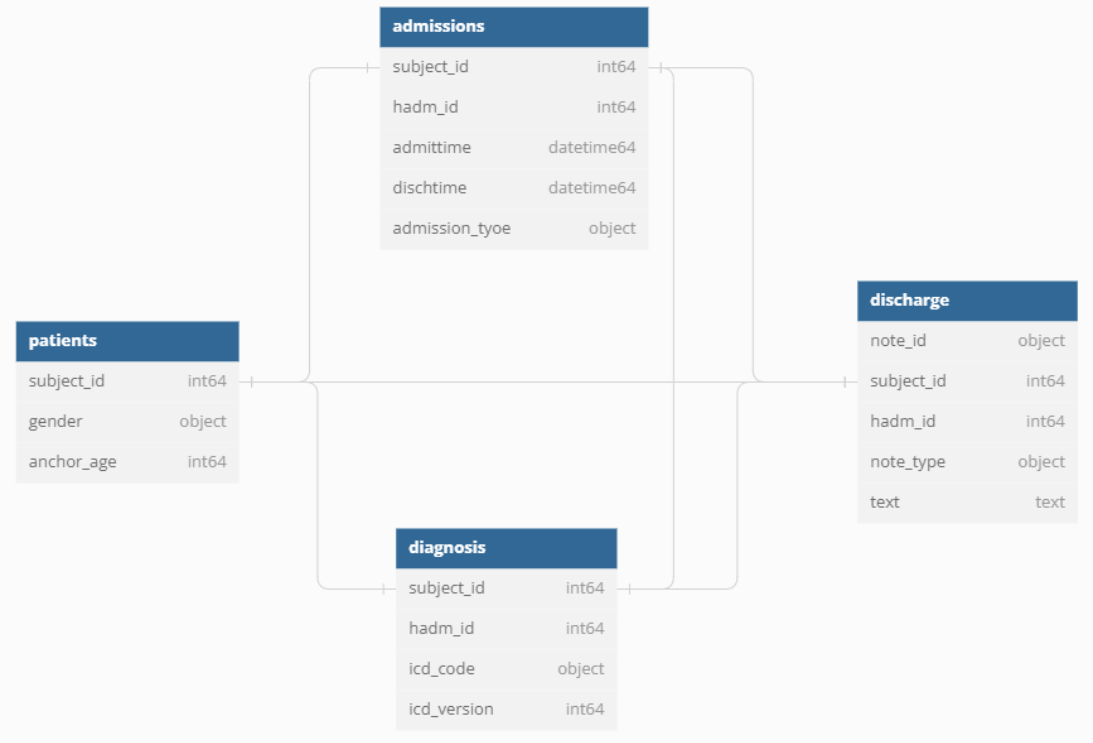

Figure 3.1 Relevant fields in each Table.

It is to note that for each of the patient (subject_id) there are one or more hospital admissions (hadm_id), and each hospital admissions are associated with one discharge note (discharge.noted_id), and each discharge note has one or more diagnoses (diagnosis.icd_code).

3.3.2. Text Cleaning

The process of de-identification and retrieval from the original EHR system introduced a number of special characters, line breaks and additional spaces (as shown in table 3.1).

Table 3.1 Example of Discharge Summary of a patient (truncated)

\nName: ___ Unit No: ___\n \nAdmission Date: ___ Discharge Date: ___\n \nDate of Birth: ___ Sex: F\n \nService: MEDICINE\n \nAllergies: \nPercocet\n \nAttending: ___.\n \nChief Complaint:\nabdominal fullness and discomfort\n \nMajor Surgical or Invasive Procedure:\n___ diagnostic paracentesis\n___ therapeutic paracentesis\n\n \nHistory of Present Illness:\n___ with HIV on HAART, COPD, HCV cirrhosis complicated by \nascites and HE admitted with abdominal distention and pain. **[continued...]**

Moreover, as notes are written by different healthcare professionals, a number of additional inconsistencies are introduced. For instance, some physicians use 'years old' while some use 'year old' to denote the age of the patient. Similarly, the terms 'female', 'f' and 'F' are used interchangeably when referring to female patients in discharge notes.

Regular Expression (Regex) is therefore used to clean the texts by removing special characters, line breaks and unwanted white spaces, and to achieve uniformity for most common adverbial phrases, adjectives, and nouns. Table 3.2 highlights two such examples of Regex to clean the texts. The full list of Regex used for this project is outlined in Appendix B.

Table 3.2 Examples of Regex to clean texts

Regex	**Objectives**
`re.sub(r'-?\byears? ?-?old\b|\by(?:o|r)*[ ./-]*o(?:ld)?\b', ' yo', text, flags=re.IGNORECASE)`	*change* 'year old', 'yearsold', etc. *to* 'yo'
`re.sub(r'\b(gentlman|male|man|m|M)(?!\S)\b', 'male', text)`	*Change* 'gentleman', 'male', 'man', 'm', 'M' *to* 'male'.

For CBOW, LSTM and CNN-LSTM models, punctuation was removed, and words were converted to lowercase. However, for attention-based models, punctuations were not removed as LLMs such as BERT to assign special tokens [CLS] at the beginning of the sentence and [SEP] at the end of the sentence to separate sentences and in the process gaining understanding of the context and boundaries of the input. Similarly, stop words were removed for conventional models using Natural Language Tool Kit (NLKT), while such step was not deemed necessary for attention-based models.

The discharge summary contains a large number of contractions (such as doesn't, won't and don't). Such contractions were expanded using python's contractions 0.1.73 library (pypi.org, 2024), as there is a lack of uniformity in the usage of contraction across MIMIC-IV discharge summaries. An example of which is illustrated in table 3.3.

Table 3.3 Example of contraction expansion

Raw format	"doesn't want to put more chemicals in her"
Fixed contraction	“does not want to put more chemicals in her”

3.3.3. Additional Variables

A number of additional were contracted from the existing variables for the purpose of this research. Examples of two additional variables is outlined in table 3.3.

Table 3.4 Additional Variables

Variable Name	Explanation	Calculation steps
days_bt_adms	“Days between admission” for a given patient.	Deducting the discharge time **dischtime** of the current hospital admission **hadm_id** from next admission time **admittime**.
next_icd_code	ICD diagnosis code for the following admission of a patient.	By shifting the **icd_code** column upwards by one row within each group of **subject_id** (patient) and **hadm_id** (admission).
next_admission_type	Admission type for the following admission of a patient.	By shifting the **admission_type** column upwards by one row within each group of **subject_id** (patient) and **hadm_id** (admission).

The calculation of days_bt_adms is particularly useful, as it allows discarding of admissions that are longer than 120 days for a patient. As explained by Guazzo et al. (2023), predicting disease in a substantially longer period becomes less useful given it lacks actionable specificity preventing the implementation of timely intervention. A number of studies including Liu et al. (2019) and Harerimana et al. (2023) used shorter timeframe for hospital readmission due to specific diseases. Harerimana et al. (2023) in their study used 90 days. However, this study used 120 days, as baseline model performance indicated similar results for 90 days and 120 days. Similarly, creation of next admission type is also important as doing so allowed this research to ignore hospital admissions that are elective, as planned and scheduled hospital admissions adversely affect predictive model performance.

3.4. ICD CODES

Given this study uses ICD codes as target variable, it is important to briefly discuss these codes. Developed by the World Health Organisation (WHO), ICD codes represent diagnoses and medical procedures in standardised way (WHO, 2019). MIMIC-IV provides two iterations of ICD codes - ICD-9 (International Classification of Diseases, 9th edition), and ICD-10 (International Classification of Diseases, 10th edition). The choice of ICD-10 iteration over previous iteration is due to improved accuracy of ICD-10 and more precisions associated with ICD-10 that make diagnoses more accurate and precise (CDC, 2015).

The first three characters of ICD-10 codes indicate the general disease or injury, which is then followed by additional characters that provide subclassifications of the disease or injury (illustrated in table 3.5).

Table 3.5 ICD-10 code for Acute Myocardial infarction.

ICD-10 code	Full diagnosis
I21	**Acute Myocardial Infarction**
I21.0	Acute transmural myocardial infarction of anterior wall
I21.1	Acute transmural myocardial infraction of inferior wall
I21.2	Acute transmural myocardial infarction of other sites
I21.3	Acute transmural myocardial infarction of unspecified site
I21.4	Acute Subendocardial myocardial infarction
I21.9	Acute myocardial infarction, unspecified.

This research study only uses the general diagnosis in line with other similar studies such as Anetta et al. (2022). ICD-10 codes were therefore truncated to the first three characters of the code. Moreover, in line with Anetta et al. (2022) and Davidson et al. (2020), the following ICD-10 categories (shown in table 3.6) were deemed to be directly related to CVD.

Table 3.6 Selected ICD-10 codes for predictions.

Diagnosis Category	General Diagnosis	ICD-10
Acute Coronary Syndrome	Myocardial Infraction	I21
	Unstable angina	I20
	Cardiac Arrest	I46
	Other Acute heart disease	I24

Heart Failure	Heart Failure	I50
Stroke	Subarachnoid haemorrhage	I60
	Intracerebral haemorrhage	I61
	Cerebral infraction	I63
	Non-specific stroke	I64
Other	Atrial fibrillation and flutter	I48
	Chronic ischemic heart disease	I25
	Nonrheumatic aortic valve disorders	I35
	Other cardiac arrhythmias	I49
	Nonrheumatic mitral valve disorders	I34

Based on the initial exploration of data, the frequency distribution of the selected ICD-10 codes is shown in figure 3.2.

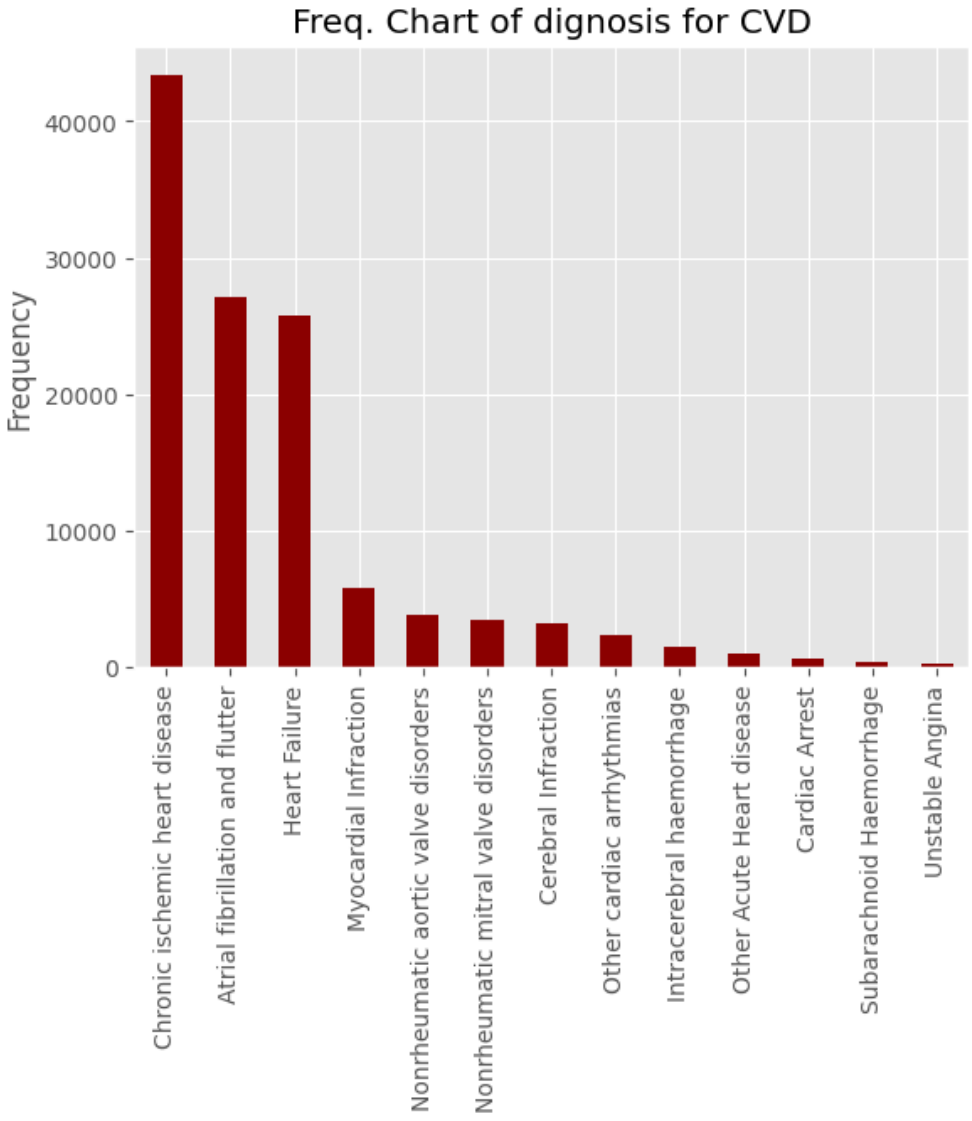

Figure 3.2 Frequency chart of CVD related ICD-10 codes.

It appears that while certain diagnoses such as Heart Failure are highly prevalent, some diagnoses such as Unstable Angina are comparatively less prevalent in the MIMIC-IV dataset. Therefore, less frequent ICD-10 diagnoses that fall into similar disease subtypes are grouped together for this research. Table 3.7 shows the grouping.

Table 3.7 Grouped ICD-10 codes for CVD.

Diagnosis	Subtype	ICD-10	Combined
Acute Coronary Syndrome	Myocardial Infraction	I21	
	Other Acute heart disease	I24	I24, I46, I20
Heart Failure	Heart Failure	I50	
Stroke	Non-specific stroke	164	160, 161, 163, 164
Other	Atrial fibrillation and flutter	I48	
	Chronic ischemic heart disease	I25	
	Nonrheumatic aortic valve disorders	I35	
	Other cardiac arrhythmias	I49	
	Nonrheumatic mitral valve disorders	I34	

3.5. SUMMARY OF DATA PREPROCESSING

The following steps (outlined in table 3.8) were taken to pre-process the data for this research.

Table 3.8 Preprocessing Steps and Rationale

Steps	Rationale
Merging data from **discharge.csv.gz**, **admissions.csv.gz** and **diagnoses_icd.csv.gz** on subject_id and hadm_id.	To obtain unified dataFrame that can be explored and filtered as required.
Separating ICD10 codes from all diagnosis codes [icd_code] and storing these codes (following modifications as shown in table 3.7) into a new column, icd_cvd.	Separation is necessary as these codes will be used as targets for the models.
Calculation of additional variables as demonstrated in table 3.4.	This was essential to filter out elective hospital admissions and admissions that are too farther in the future.

Cleaning of discharge summaries using Regex and additional libraries for all models.	This was necessary to remove noise, typos, errors and create uniform representation of the texts for the models.
Removing any records with blank discharge summary and diagnosis code.	As these contain incomplete information or noise that would have impacted model performance.
Removing discharge notes for last admission of each patient.	Since, this study uses discharge notes of time t_n to predict CVD diagnosis code for $t_{n+1,}$ notes from t_{n+1} therefore do not add any value.

3.5.1. Targets

As mentioned, this research aims to predict CVDs among patients. This is done by predicting diagnosis code related to CVD for the next hospital visit. This therefore makes the approach a Multilabel Classification task, where 9 ICD codes of CVDs (table 3.7) are to be classified (either 1 or 0) by the model. While initially, 9 labels were identified as targets, however, subsequent filtering of data (as outlined in table 3.8) resulted in significantly unbalanced dataset, where top three CVD diagnostic codes (I50, I25 and I48) account for over 82% of the diagnostic codes (as illustrated in figure 3.3).

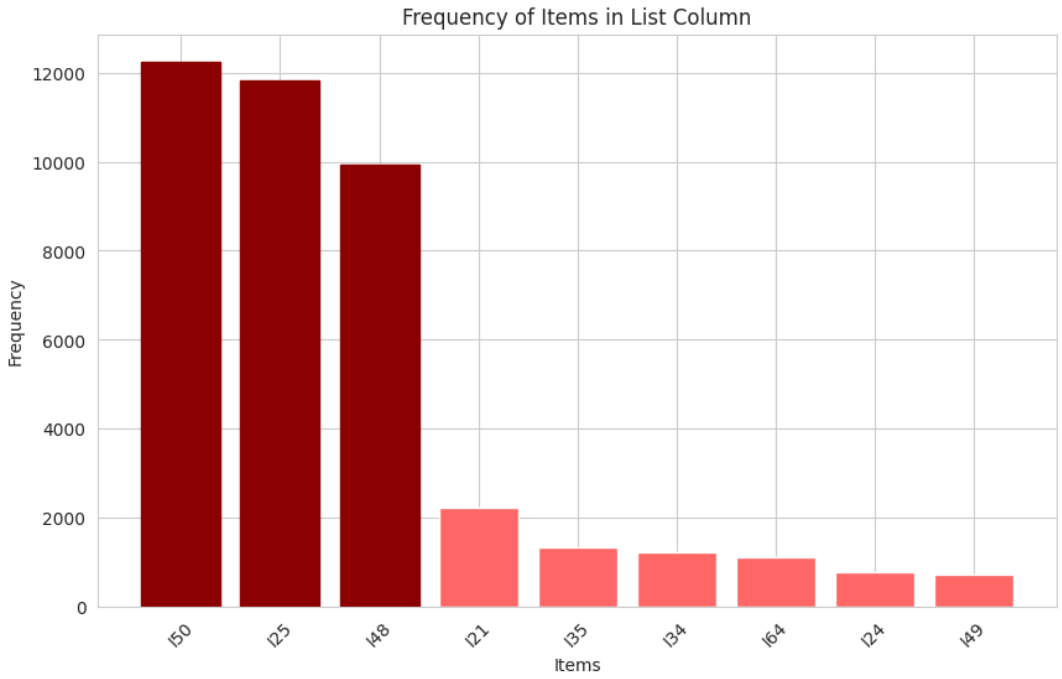

Figure 3.3 Frequency of CVD ICD codes in the final dataset.

Furthermore, steps such as under-sampling of majority classes (I50, I25 and I48) significantly reduced data available for training the models. Over-sampling of medical text is strongly discouraged, as over-sampling and other similar practices can affect real-world medical applications (Alkhawaldeh et al. 2023). Therefore, this study decided to only include three majority classes in the classification. This is further justified by the mortality rates associated with these diagnostic codes. The final dataset selected for training and evaluating models consists of 14,181 patient records of whom 1,173 (8.3%) patients were deceased. For 69% of the deceased patients, CVD disease was primary or secondary cause of death. Furthermore, as shown in figure 3.4, the majority of the fatality from CVD diseases can be contributed to the 3 majority classes. This therefore indicates high prevalence of these CVD diseases, as well as high morality associated with these diseases. Focus on these three CVD disease classes are therefore justified.

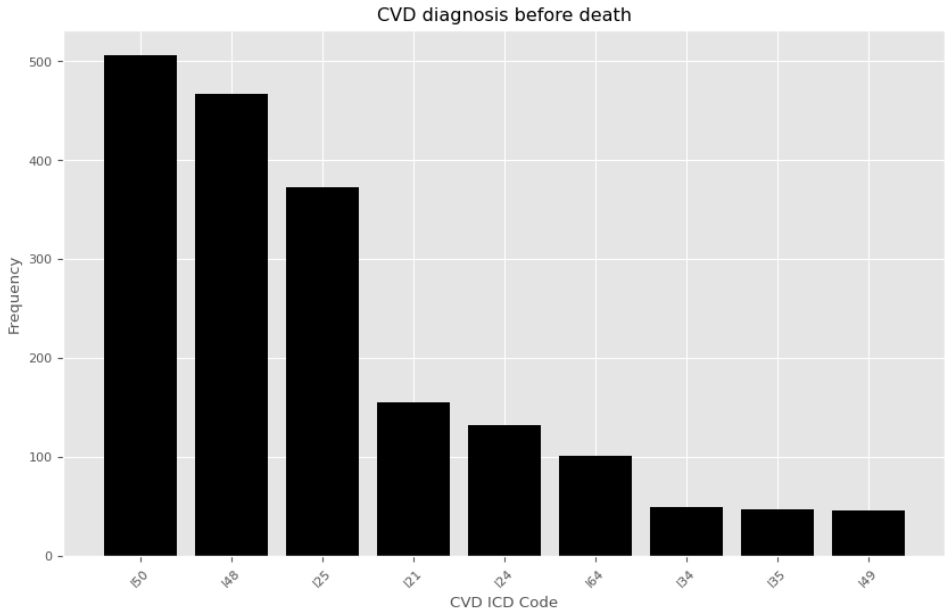

Figure 3.4 Prevalence of CVD ICD Code in deceased patients.

3.6. Evaluation Metrics

Based on similar studies in the established literature, this study adopts widely used performance metrics for evaluating the models assessed. These include macro-averaged and micro-averaged F1, Accuracy, Area Under the Curve (AUC), Precision and Recall.

3.6.1. Accuracy

Accuracy is a widely used metric that indicates the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances (Hicks et al., 2022). This metric illustrates the overall correctness of a model's predictions.

This is calculated using the formula shown in equation 3.1.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Equation 3.1 formula for calculating Accuracy.

As Hicks et al. (2022) explain, this metric is simple and provides a quick and intuitive measure of a model's ability to make correct predictions overall. However, accuracy has some limitations, especially for imbalanced datasets. For such datasets, a model can simply predict the majority class and achieve high accuracy while ignoring the minority classes (Ochella & Shafiee, 2021). Therefore, this metric should be examined along with other metrics.

3.6.2. Precision

This metric measures the ratio of True Positive (model correctly predicting the positive class) predictions to the total number of predictions that include False Positive (model incorrectly predicting the positive class) and True Positive (Ochella & Shafiee, 2021). Equation 3.2 shows the formula to calculate precision.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Equation 3.2 formula for Precision

3.6.3. Recall

This metric measures the ratio of True Positive predictions to the total number of actual predictions that include False Negatives (model incorrectly predicting the negative class) and True Positive (Hicks et al., 2022). The formula for recall is shown in equation 3.3.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Equation 3.3 formula for Recall

3.6.4. F1-Micro and F1-Macro

This metric computes the global average F1-score by aggregating the contribution of all classes. According to Hicks et al. (2022), this metric treats each class equality, while giving more weight to more frequent classes, which helps to measure overall accuracy of the model.

On the other hand, F1-Macro calculates the performance of each class independently and then takes the average, and each class regardless of their frequency receives equal weights (Hicks et al., 2022).

3.6.5. Area Under the Curve (AUC)

This measure measures the ability of a model to distinguish between classes. This Metric provides insight into a model's ability to differentiate between positive and negative cases (Ochella & Shafiee, 2021). This is calculated (as shown in equation 3.4) from the ROC curve using trapezoidal rule, which sums the area of trapezoids that are under the curve (ROC curve). If (x_1, y_1), (x_2, y_2) …, (x_n, y_n) are points under the ROC curve, then AUC is given by,

$$AUC = \sum_{i=1}^{n-1} \frac{(x_{i+1} - x_i) \cdot (y_i + y_{i+1})}{2}$$

Equation 3.4 Formula for AUC

3.7. HARDWARE

For initial data cleaning, exploration and assessing baseline DL models were performed on a personal computer with AMD Ryzen R7-5700x CPU with 96 GB of RAM and NVIDIA RTX 3060 (12 GB GDDR6). However, larger models such as clinical-Longformer, Hi-HBERT and BERT (with sliding window) were run on Google Colab using either A100 GPU (System RAM: 83.5 GB, GPU RAM: 40.0 GB) or T4 GPU (System RAM: 51.0 GB, GPU RAM: 15.0 GB) depending on the model needs.

4. EMPIRICAL STUDY

4.1. CHAPTER INTRODUCTION

The purpose of this chapter is to present the demographic characteristics of the patients in the MIMIC-IV dataset, followed by discussion of performance of various models assessed for this study. Finally, this chapter also presents and discusses the recommended model to predict CVD for patients from their discharge summaries.

4.2. DEMOGRAPHIC INFORMATION

According to Turchioe et al. (2022), discussion of demographic characteristics of the participants helps to determine whether the findings and any proposed model can be generalized. Moreover, discussion of demographic information allows other researchers to reproduce the findings as well as replicate model performance.

MIMIC-IV dataset contains medical information on 180,733 patients that attended Emergency Departments (excluding ICU) of Beth Israel Deaconess Medical Center between 2008 and 2019. The demographic distributions of the patients are shown in table 4.1.

Table 4.1 Demographic Information of the participants in the Dataset

Demographic Category	Sub-category	Frequency
Gender	Male	85,004 (47.03%)
	Female	95,729 (52.97%)
Race	White	120,853 (66.87%)
	Black	23,409 (12.95%)
	Hispanic/Latino	9,765 (5.40%)
	Asian	7,522 (4.16%)
	Other	19,184 (10.61%)
Marital Status	Married	77,815 (43.06%)
	Single	67,500 (37.38%)
	Widowed	16,728 (9.26%)
	Divorced	11,370 (6.29%)
	Unknown	7,270 (4.02%)

The age distribution of the patients in the database is shown in figure 4.1. It is to note that large concentrations on 91 year is due to setting the age of all patients aged over 89 as 91-year-old in line with the US legislation governing patient privacy (Johnson et al., 2023).

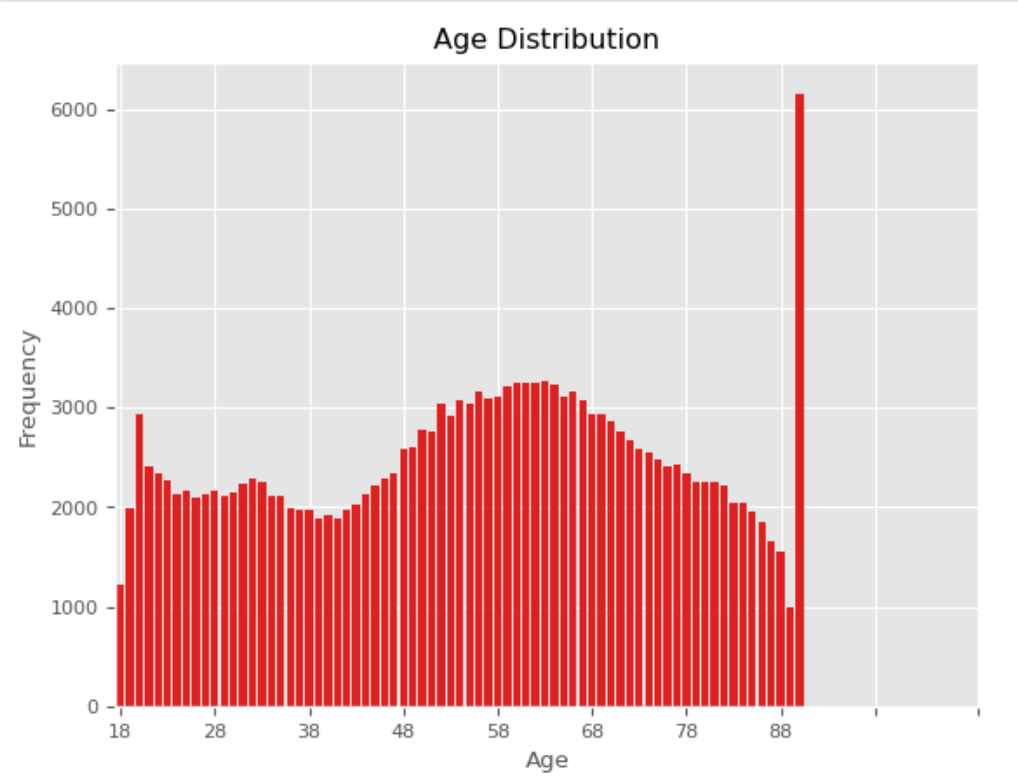

Figure 4.1 Age distribution of the patients in the dataset.

4.3. Models Selection

Although the focus on this study is attention-based models, widely used conventional DL models are initially assessed to establish baseline performance. Furthermore, Zhan et al. (2021) and Rishivardhan et al. (2020) in their study concluded that in many NLP tasks in clinical settings, attention-based models tend to underperform conventional ML methods. Therefore, the inclusion of conventional models would help to determine whether such a claim holds. Besides,

The first model assessed in this research is Bidirectional LSTM adopted in the study of Guazzo et al. (2023) to predict hospitalization of diabetes patients due to CVD using medical notes as input. Given this research aims to predict CVD diagnosis of patients in the near future, the evaluation of this model is therefore warranted. This model further acts as a benchmark to assess other models.

As mentioned, BERT is one of the most widely used attention-based models for various downstream tasks including Multilabel classification, as the model is already pre-trained on large corpora on vast amount of textual data. Furthermore, as Houssein et al. (2023) in their study found that BERT along with CharacterBERT outperformed DL models in identifying heart disease risk factors from clinical notes.

ClinicalBERT, which is a variant of BERT is also assessed as this attention-based model is fine-tuned on clinical text data to enhance its performance in NLP tasks related to healthcare (Turchin et al., 2023). Moreover, this model was specifically trained on large corpora of EHRs, which makes it more adept to understanding medical abbreviations, terminology, and complex and unique linguistic patterns that emerge in unstructured clinical notes. Therefore, this model is assessed in this study to determine whether pre-trained on similar data can help to enhance model performance.

This study also included ConBERT, which is similar attention-based model adopted by Houssein et al. (2023), which integrates BERT with CharacterBERT (illustrated in figure 4.2) to improve performance as this model is able to capture rich lexical information through BERT while capturing morphological information through CharacterBERT.

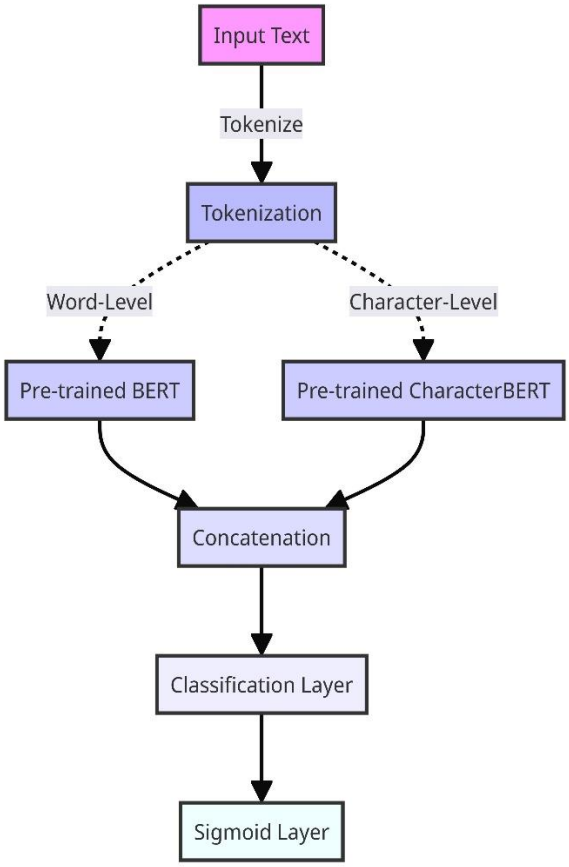

Figure 4.2 ConBERT Architecture

While the models thus far just focus on unstructured clinical and/or medical notes, a number of recent studies attempted to incorporate time progression in disease predictions, as naturally diseases often follow specific progression patterns, and therefore incorporation of time progression helps models to make better predictions (Nia et al., 2023). Moreover, as explained by Li et al. (2020), EHRs contain vast amounts of temporal data, and effectively leveraging these data helps to enhance predictive power of diagnostic models. Therefore, Li et al. (2020) proposed BHERT model that analyzes temporal sequences of medical events in EHRs to predict diseases. In line with BHERT model proposed by Li et al. (2020), this study also examined modified BHERT ($BHERT_{mod}$) model (figure 4.3) to determine whether incorporation of temporal data (in the form of age progression and duration between admission) improve CVD diagnosis prediction.

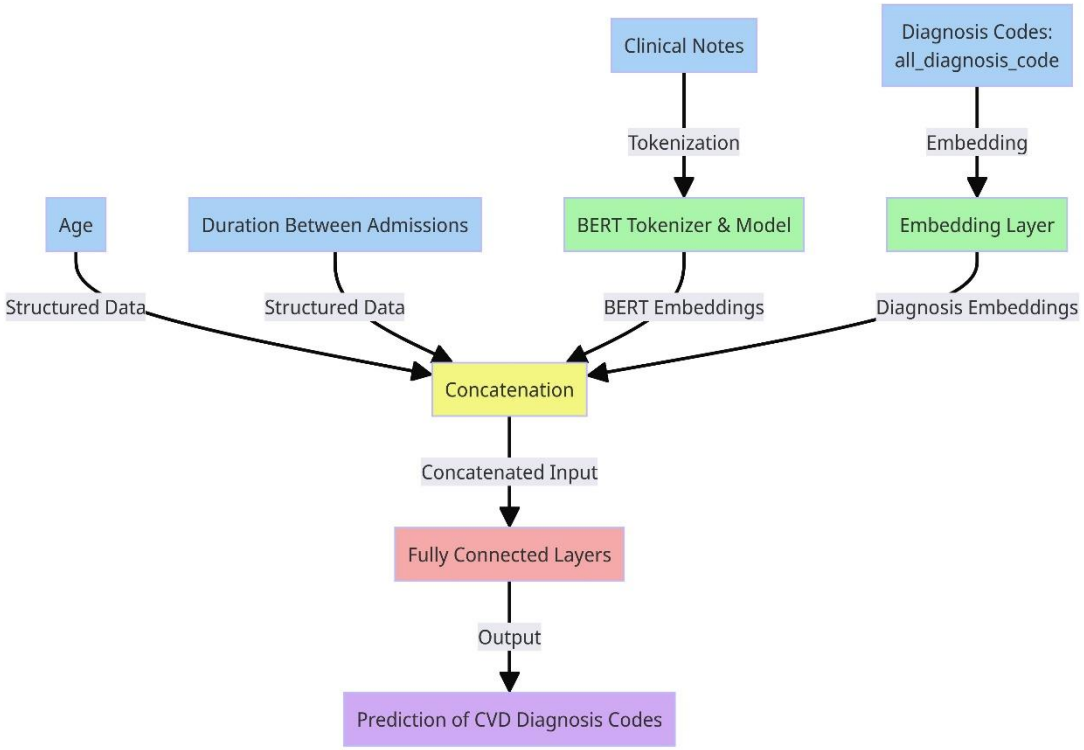

Figure 4.3 $BHERT_{mod}$ model architecture.

One of the key limitations of BERT and clinicalBERT is that these are limited as they can only handle input sequences up to 512 tokens in length (Turchin et al., 2023). This poses a significant challenge in healthcare, as many clinical notes, patient documents and unstructured notes can be significantly long, which would result in BERT and clinicalBERT failing to capture important information from larger documents. Clinical Longformer aims to address these limitations as this model is adept in handling long documents more efficiently. By abandoning self-attention mechanisms of BERT, and embracing local and global attention

patterns, Clinical Longformer is able to process length EHR records, which therefore makes it an ideal model to achieve the aim of this research.

However, recently, a new approach has been adopted to address the limitation of maximum input size of BERT. This is known as a Sliding window approach. This method involves breaking down a long text into smaller, overlapping segments known as windows, which are then individually processed by BERT (illustrated in figure 4.4). BERT then aggregates the results from these windows to make predictions.

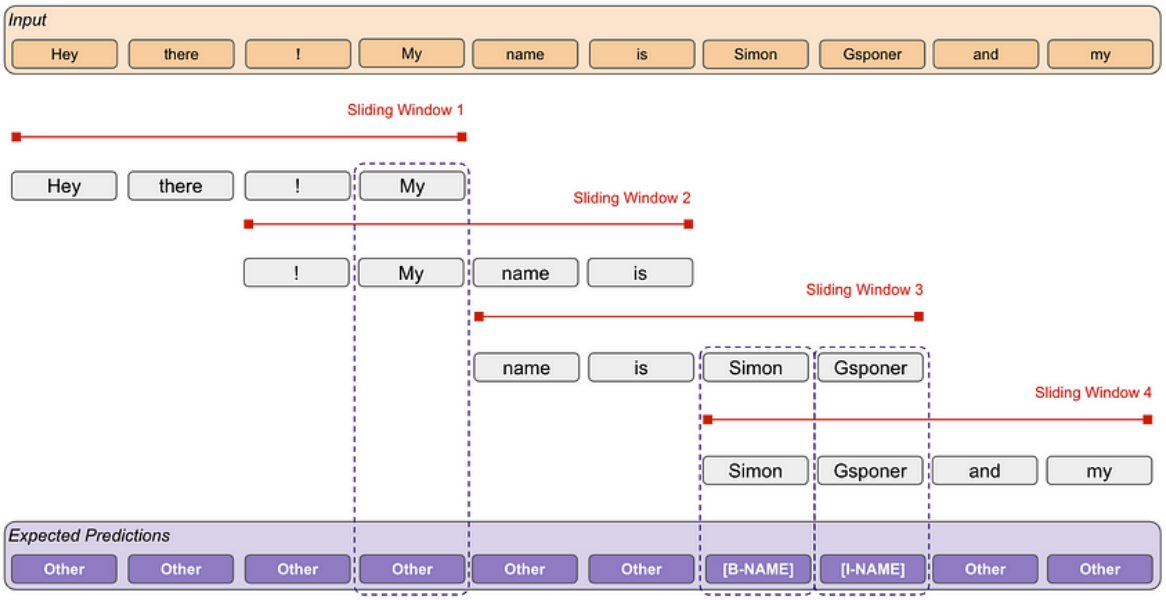

Figure 4.4 Examples of sliding window used in BERT (Gsponer, 2024).

However, this research also proposes a novel approach that is being widely adopted in Retrieval Augmented Generation (RAG). The strategy is known as “chunking”, which simply refers to breaking down large texts into smaller texts to optimize relevancy (Schwaber-Cohen, 2024). There are different chunking strategies. However, this research adopts semantic chunking, which refers to creating chunks based on meaning of segments within the document (Schwaber-Cohen, 2024). The process of chunking undertaken for the proposed model is shown in figure 4.5.

As it appears, discharge note from each admission was split into chunks (using **Langchain’s** RecursiveCharacterTextSplitter). Relevant text samples are then defined using CVD risk factors that were identified in the literature review (table 2.3), as well as from the study of Houssein et al. (2023). The text samples used for semantic Chunking is shown in table 4.2.

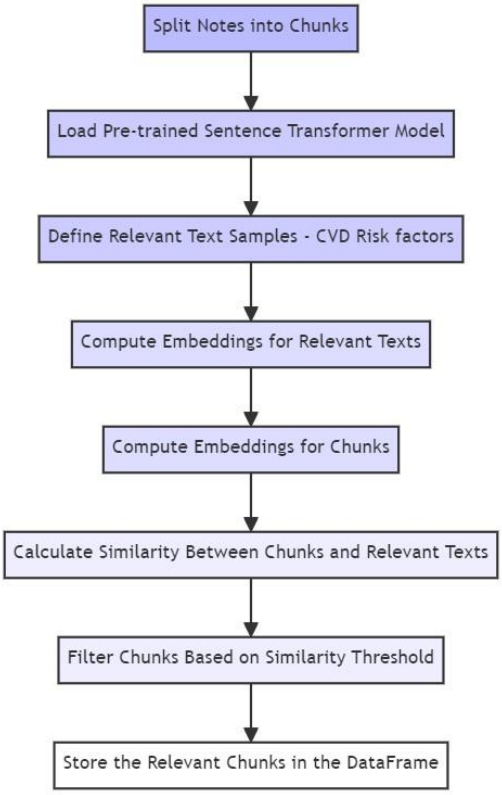

Figure 4.5 Semantic Chunking Process.

Once the relevant text samples are defined, embeddings of these texts are then calculated along with embeddings of the chunks. Based on cosine similarities between the chunks and relevant texts, relevant chunks are filtered and stored in the Data Frame.

Table 4.2 Relevant Texts used for Semantic Chunking

Text Samples related to CVD Risk Factors
["smoking", "drinking", "heart", "blood pressure", "BMI", "medication", "male", "hypertension", "physical activity", "diabetes", "obesity", "overweight", "lipoprotein cholesterol", "diet", "grip strength", "family history", "pollution", "diabetes"]

Relevant Chunks were given priority. Relevant chunks along with other chunks are then inputted into BERT model to predict CVD diagnosis of the patients. The model architecture of $BERT_{chunking}$ is shown in figure 4.6.

Similarly, $ClinicalBERT_{chunking}$ using the same chunks and ClinicalBERT model was also assessed.

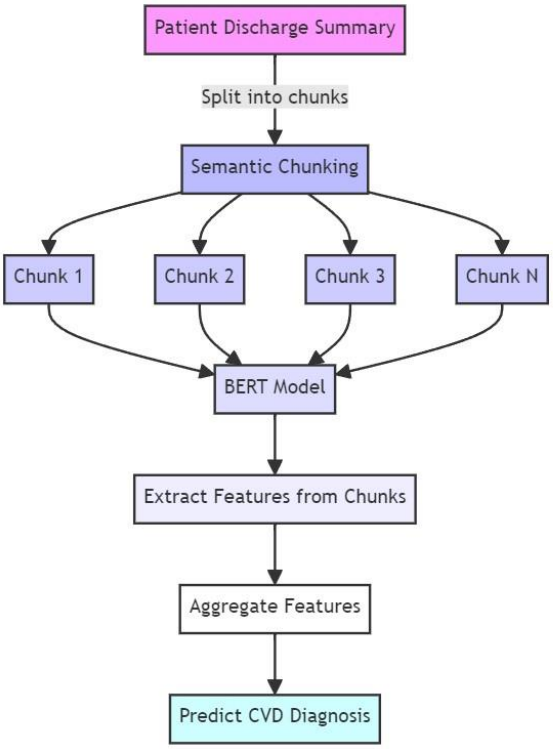

Figure 4.6 $BERT_{chunking}$

4.4. MODEL EVALUATION

Table 4.3 shows summary of the model performances that are described in the previous section.

Table 4.3 Performance of different models

Models	Accuracy	AUC	F1-Micro	F1-Macro
LSTM	0.18	0.5	0.0	0.0
$BERT_{base}$	0.56	0.78	0.54	0.23
$BHERT_{mod}$	0.57	0.79	0.51	0.17
ConBERT	0.57	0.69	0.53	0.21
Clinical-LongFormer	0.49	0.60	0.53	0.11
clinicalBERT	0.57	0.73	0.69	0.25
$BERT_{sliding\_window}$	**0.73**	**0.83**	**0.80**	0.51
$BERT_{chunking}$	0.71	0.79	0.73	**0.72**
$ClinicalBERT_{chunking}$	0.70	0.80	0.73	**0.72**

From table 4.3, it appears that $BERT_{sliding\_window}$ is the best model, as it has the highest accuracy, AUC and F1-Micro scores. This indicates that $BERT_{sliding\_window}$ outperformed all the models in terms of overall prediction accuracy as well as discriminative power (AUC). This is followed by $BERT_{chunking}$ and $clinicalBERT_{chunking}$, as both of these models performed similarly. The following chapter provides the discussion of the performance and explanation for underperformance of other models such as $BHERT_{mod}$ and Clinical-LongFormer.

5. RESULTS AND DISCUSSION

5.1. CHAPTER INTRODUCTION

The previous chapter presented the model assessed for this research, as well as their performance. This chapter discusses those performance and provides recommendations for the suitable model for CVD predictions in patients. Furthermore, comparative analysis is also conducted with similar studies from the existing literature.

5.2. DISCUSSION OF MODELS PERFORMANCE

As presented in table 4.3, LSTM performed poorly across all performance metrics. LSTM and other DL models are less effective than attention-based models for processing complex and long text data, such as the discharge notes used in this study. Long-range dependencies and contexts that are hallmark of discharge notes are key challenges or LSTM. Therefore, for longer medical texts, LSTM is unlikely to perform well compared to attention-based models.

The base model of BERT has shown significant improvement over LSTM. This could be contributed to BERT's ability to capture complex relationships and contextual information within the discharge notes. However, considering base model of BERT was not fine-tuned for clinical texts, and the limitations of 512 likely resulted in comparatively poor performance of the model. However, surprisingly, clinicalBERT despite fined tuned for clinical text performed somewhat similar to the base model of BERT. This could be contributed to the input limitations of clinicalBERT as highlighted by Turchin et al. (2023).

The performance of ConBERT also shows that while this model was able to capture some aspects of the data, it struggled to discriminate between positive and negative cases (indicated by lower AUC). Similarly, Clinical-LongFormer also performed poorly. Despite being designed to handle longer sequences; the poor performance could be attributed to the model's failure to handle complexity in the clinical text. Lower F1-Macro indicates that this model struggled to classify less frequent classes correctly.

The performance of $BHERT_{mod}$ was similar to the performance of BERT. BHERT model was developed for EHR data from GP practices in the UK. Such data contain longitudinal information of patients with temporal context, while discharge notes of the patients used in this study focus on specific hospital admission of a patient with only relevant acknowledgement of medical information to that particular hospital stay.

Finally, as mentioned, $BERT_{sliding\_window}$ is the best model, as it has the highest accuracy, AUC and F1-Micro scores as it is able to efficiently handle large texts by preserving contexts against segments, which therefore led to higher accuracy and AUC. Similar chunking enabled BERT to prioritize nuanced and dense information from the discharge notes, which therefore led to

good model performance. Similarly, clinicalBERT$_{chunking}$ slightly outperformed BERT$_{chunking}$, which could be attributed to ClinicalBERT's ability to understand and process clinical texts better.

5.3. COMPARATIVE ANALYSIS

Comparative analysis is difficult to conduct. This is because as discussed, there is a lack of studies that used MIMIC-IV as input data given the data was made public regularly. Furthermore, another key difference between this study and large number of studies in the existing literature is that while this study attempted to predict CVD for patients in the near future (within 120 days), other studies are more concerned with identifying diagnosis of CVD diseases from the medical notes.

However, Liu et al. (2019) in their study used MIMIC-III, previous iteration of the data used in this study to predict hospital readmission. Using CNN on the discharge summaries, the authors managed to achieve F1-score of 0.733 for 30-day readmission, and 0.756 for general readmission. However, BERT$_{sliding_window}$ model of this study outperformed (F1-Micro of 0.80) the model of Liu et al. (2019), which indicates transformer-based models' ability to capture more nuanced patterns in medical texts compared to CNNs.

However, the model developed by Qiu et al. (2022) outperformed all the models of this study, as using Risk Factor Attention-based Model (RFAB) with BiLSTM-CRF, the authors were able to achieve an F1-score of 0.9586, which demonstrates strong performance. However, without access to data used by the authors, it is difficult to ascertain whether the same model would have performed similarly with MIMIC-IV discharge notes.

Sliding window and semantic chunking models of this study have outperformed the LSTM model used by Guazzo et al. (2023) in predicting hospitalization of diabetes patients from their medical reports. While for longer period of time, their model performed better, for shorter window, such as 6-months readmissions, their model struggled and only achieved F-1 score of 0.64, while F1-score of best performing model of this study achieved 0.80 in predicting hospital attendance of patients with specific CVD diagnosis. This therefore indicates superiority of transformer-based models over conventional DL models.

Two of the studies included in the literature review focused on extracting medical information from unstructured data. Adekkannattu et al. (2019) used EchoExtractor, a purpose-built NLP tool for extracting key clinical events, while Viani et al. (2019) used RNNs to achieve the same goal albeit from different dataset. Both studies achieved F-1 score > 0.90. While this study did not measure the efficacy of the medical information extractor as part of semantic chunking, it appears that the extracted information significantly improved the performance of the BERT model, which indicates the efficacy of semantic chunking in capturing key medical events from unstructured data.

6. CONCLUSIONS AND FUTURE WORKS

6.1. SUMMARY OF THE FINDINGS

CVD remains the leading global cause of death contributing significantly to mortality and morbidity across different demographics. This study aimed to leverage attention-based NLP models to predict severe forms of CVD from unstructured clinical notes. Through a comparative analysis of various models, including LSTM, BERT, ClinicalBERT, and their variants with sliding window and chunking approaches, the study identified that transformer-based models, particularly BERT with sliding window as well as semantic chunking, outperform traditional deep learning models in handling complex, long-text data from discharge summaries. The best-performing model, BERT (sliding window), demonstrated the highest Accuracy, AUC, and F1-Micro scores, indicating its efficacy in predicting CVD from unstructured clinical notes. This research therefore highlights the potential of advanced NLP models in making clinical predictions, which then enabled timely interventions.

6.2. LIMITATIONS

However, this study has some limitations. The first limitation is Data Specificity since this study solely relied on discharge summaries from MIMIC-IV database, which may not generalize to other health settings or medical records. Mostly importantly while the models performed well, the absence of temporal dynamics might have limited their predictive power as additional contexts were missed. The fourth limitation is data modality as highlighted by several studies in literature including Anette et al. (2022). Due to data limitations, several CVD disease predictions were abandoned. Furthermore, this study uses train-test split for evaluating the model. Although cross-validation is more suitable to reduce variance, better detect overfitting and other benefits, due to high computational costs associated with cross-validation, a single train-test split was performed.

Another key limitation of this study is comparative analysis constraints. Besides the research objectives and model architectures, significant differences in data used by different studies in the literature make comparative analysis highly challenging. Data confidentiality required to study clinical notes from EHR makes it more challenging to fully understand and compare datasets used by different authors, which then makes transparency and reproducibility challenging. Finally, computational resources required to train attention-based models further limited wider exploration of models and larger embedding, which could have further improved performance of the models in this study.

6.3. FUTURE WORK

Future studies should explore how to integrate temporal information such as progression of CVD risk factors, symptoms, and comorbidities over time to enhance model performance. Besides relying on EHR data from hospital, efforts should be made to incorporate data from GP practices to obtain more longitudinal data that can help with temporal dynamics. While the purpose of this study is to focus on unstructured data due to resource limitations and time constraints, future studies should explore whether incorporating structured data could improve the performance of the models identified in this study. Finally, future studies should explore implementation of real-time clinical decision support systems that can equip healthcare professionals to make timely interventions.

BIBLIOGRAPHICAL REFERENCES

Abrahão, M. T., Nobre, M. R., & Gutierrez, M. A. (2017). A method for cohort selection of Cardiovascular Disease Records from an electronic health record system. *International Journal of Medical Informatics*, *102*, 138–149. https://doi.org/10.1016/j.ijmedinf.2017.03.015

Alkhawaldeh, I. M., Albalkhi, I., & Naswhan, A. J. (2023). Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, *13*(5), 373–378. https://doi.org/10.5662/wjm.v13.i5.373

Altarriba, J., & Basnight-Brown, D. (2022). The Psychology of Communication: The interplay between language and culture through time. *Journal of Cross-Cultural Psychology*, *53*(7–8), 860–874. https://doi.org/10.1177/00220221221114046

Althubaiti, A. (2016). Information bias in health research: Definition, Pitfalls, and Adjustment Methods. *Journal of Multidisciplinary Healthcare*, 211. https://doi.org/10.2147/jmdh.s104807

Anetta, K., Horak, A., Wojakowski, W., Wita, K., & Jadczyk, T. (2022). Deep Learning Analysis of Polish electronic health records for diagnosis prediction in patients with cardiovascular diseases. *Journal of Personalized Medicine*, *12*(6), 869. https://doi.org/10.3390/jpm12060869

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR 2015*. https://doi.org/10.48550/arXiv.1409.0473

Bays, H. E., Taub, P. R., Epstein, E., Michos, E. D., Ferraro, R. A., Bailey, A. L., Kelli, H. M., Ferdinand, K. C., Echols, M. R., Weintraub, H., Bostrom, J., Johnson, H. M., Hoppe, K. K., Shapiro, M. D., German, C. A., Virani, S. S., Hussain, A., Ballantyne, C. M., Agha, A. M., & Toth, P. P. (2021). Ten things to know about ten cardiovascular disease risk factors. *American Journal of Preventive Cardiology*, *5*, 100149. https://doi.org/10.1016/j.ajpc.2021.100149

Boren, S., & Moxley, D. (2015). Systematically Reviewing the Literature: Building the Evidence for Health Care Quality. *Mo Med*, *112*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6170102/

Cassar, A., Holmes, D. R., Rihal, C. S., & Gersh, B. J. (2009). Chronic coronary artery disease: Diagnosis and management. *Mayo Clinic Proceedings*, *84*(12), 1130–1146. https://doi.org/10.4065/mcp.2009.0391

CDC. (2023, May 15). Heart disease facts. https://www.cdc.gov/heartdisease/facts.htm

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support Vector Machine Classification: Applications, challenges and Trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Chai, C. P. (2022). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/s1351324922000213

Chen, S.-F., Loguercio, S., Chen, K.-Y., Lee, S. E., Park, J.-B., Liu, S., Sadaei, H. J., & Torkamani, A. (2023). Artificial Intelligence for risk assessment on primary prevention of coronary artery disease. *Current Cardiovascular Risk Reports*, *17*(12), 215–231. https://doi.org/10.1007/s12170-023-00731-4

Chervonenkis, A. Ya. (2013). Early history of Support Vector Machines. *Empirical Inference*, 13–20. https://doi.org/10.1007/978-3-642-41136-6_3

Chokwijitkul, T., Nguyen, A., Hassanzadeh, H., & Perez, S. (2018). Identifying risk factors for heart disease in electronic medical records: A deep learning approach. *Proceedings of the BioNLP 2018 Workshop*. https://doi.org/10.18653/v1/w18-2303

Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of Artificial Intelligence*, 603–649. https://doi.org/10.1007/978-81-322-3972-7_19

Cosselman, K. E., Navas-Acien, A., & Kaufman, J. D. (2015). Environmental factors in cardiovascular disease. *Nature Reviews Cardiology*, *12*(11), 627–642. https://doi.org/10.1038/nrcardio.2015.152

Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F., & Zalewski, A. (2017). Electronic Health Records to facilitate clinical research. *Clinical Research in Cardiology*, *106*(1), 1–9. https://doi.org/10.1007/s00392-016-1025-6

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Proceedings of the 2019 Conference of the North*. https://doi.org/10.18653/v1/n19-1423

Elkin, P. L., Mullin, S., Mardekian, J., Crowner, C., Sakilay, S., Sinha, S., Brady, G., Wright, M., Nolen, K., Trainer, J., Koppel, R., Schlegel, D., Kaushik, S., Zhao, J., Song, B., & Anand, E. (2021). Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *Journal of Medical Internet Research*, *23*(11). https://doi.org/10.2196/28946

Farrag, A., Eraky, A., Aroussy, W., Sayed, G., Mahrous, A., Adel, A., Ibrahim, A., Ibrahim, A., & M. (2015). Obesity and other cardiovascular risk factors in Egyptian university students: Magnitude of the problem. *Epidemiology: Open Access*, *05*(01). https://doi.org/10.4172/2161-1165.1000181

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, *279*(1747), 4522–4531. https://doi.org/10.1098/rspb.2012.1741

Górriz, J. M., Álvarez-Illán, I., Álvarez-Marquina, A., Arco, J. E., Atzmueller, M., Ballarini, F., Barakova, E., Bologna, G., Bonomini, P., Castellanos-Dominguez, G., Castillo-Barnes, D., Cho, S. B., Contreras, R., Cuadra, J. M., Domínguez, E., Domínguez-Mateos, F., Duro, R. J., Elizondo, D., Fernández-Caballero, A., … Ferrández-Vicente, J. M. (2023). Computational approaches to explainable artificial intelligence: Advances in theory, applications and Trends. *Information Fusion*, *100*, 101945. https://doi.org/10.1016/j.inffus.2023.101945

Gsponer, S. (2024, February 10). *A comprehensive guide: Using a bert LLM on texts exceeding the maximum input size*. Medium. https://medium.com/@simon.gsponer/a-comprehensive-guide-using-a-bert-llm-on-texts-exceeding-the-maximum-input-size-47d1b72e397f

Guazzo, A., Longato, E., Fadini, G. P., Morieri, M. L., Sparacino, G., & Di Camillo, B. (2023). Deep-learning-based natural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visits' text. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-45115-1

Harerimana, G., Kim, G. I., Kim, J. W., & Jang, B. (2023). HSGA: A hybrid LSTM-CNN self-guided attention to predict the future diagnosis from discharge narratives. *IEEE Access*, *11*, 106334–106346. https://doi.org/10.1109/access.2023.3320179

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of Artificial Intelligence. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-09954-8

Holmes, J. H., Beinlich, J., Boland, M. R., Bowles, K. H., Chen, Y., Cook, T. S., Demiris, G., Draugelis, M., Fluharty, L., Gabriel, P. E., Grundmeier, R., Hanson, C. W., Herman, D. S., Himes, B. E., Hubbard, R. A., Kahn, C. E., Kim, D., Koppel, R., Long, Q., … Moore, J. H. (2021). Why is the electronic health record so challenging for research and clinical care? *Methods of Information in Medicine*, *60*(01/02), 032–048. https://doi.org/10.1055/s-0041-1731784

Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and Deep Learning Techniques. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-34294-6

Hu, S.-Y., Santus, E., Forsyth, A. W., Malhotra, D., Haimson, J., Chatterjee, N. A., Kramer, D. B., Barzilay, R., Tulsky, J. A., & Lindvall, C. (2019). Can machine learning improve patient selection for cardiac resynchronization therapy? *PLOS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0222397

Instituto Nacional de Estatistica. (2023). *In 2021, there was an increase in deaths from malignant neoplasms of the trachea, bronchi and lung, besides those of COVID-19 - 2021*. Statistics portugal - web portal. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=594418921&DESTAQUESmodo=2

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). Author correction: Mimic-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1). https://doi.org/10.1038/s41597-023-01945-2

Jones, K. S. (1994). Natural language processing: A historical review. *Current Issues in Computational Linguistics: In Honour of Don Walker*, 3–16. https://doi.org/10.1007/978-0-585-35958-8_1

Joopudi, V., Dandala, B., & Devarakonda, M. (2018). A convolutional route to abbreviation disambiguation in clinical text. *Journal of Biomedical Informatics*, *86*, 71–78. https://doi.org/10.1016/j.jbi.2018.07.025

Joseph, P., Leong, D., McKee, M., Anand, S. S., Schwalm, J.-D., Teo, K., Mente, A., & Yusuf, S. (2017). Reducing the global burden of cardiovascular disease, part 1. *Circulation Research*, *121*(6), 677–694. https://doi.org/10.1161/circresaha.117.308903

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Koene, R. J., Prizment, A. E., Blaes, A., & Konety, S. H. (2016). Shared risk factors in cardiovascular disease and cancer. *Circulation*, *133*(11), 1104–1114. https://doi.org/10.1161/circulationaha.115.020406

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. https://doi.org/10.1073/pnas.97.22.11850

Lasnik, H., & Lohndal, T. (2017). Noam Chomsky. *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.356

Lefkowitz, M. (2019, September 25). *Professor's Perceptron paved the way for AI – 60 years too soon*. Cornell Chronicle. https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in Natural Language Processing: A Survey. *AI Open*, *3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., & Rahimi, K. (2023). Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*, *27*(2), 1106–1117. https://doi.org/10.1109/jbhi.2022.3224727

Li, Y., Rao, S., Solares, J. R., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). Behrt: Transformer for Electronic Health Records. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-62922-y

Liao, W., & Voldman, J. (2023). A multidatabase extraction pipeline (metre) for facile cross validation in Critical Care Research. *Journal of Biomedical Informatics*, *141*, 104356. https://doi.org/10.1016/j.jbi.2023.104356

Liu, X., Chen, Y., Bae, J., Li, H., Johnston, J., & Sanger, T. (2019). Predicting heart failure readmission from clinical notes using deep learning. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/bibm47256.2019.8983095

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 471–484. https://arxiv.org/pdf/1907.11692.pdf

Locati, E. T., Van Dam, P. M., Ciconte, G., Heilbron, F., Boonstra, M., Vicedomini, G., Micaglio, E., Ćalović, Ž., Anastasia, L., Santinelli, V., & Pappone, C. (2023). Electrocardiographic temporo-spatial assessment of depolarization and repolarization changes after epicardial arrhythmogenic substrate ablation in Brugada syndrome. *European Heart Journal - Digital Health*, *4*(6), 473–487. https://doi.org/10.1093/ehjdh/ztad050

Lopez, E. O., Ballard, B. D., & Jan, A. (2023). *Cardiovascular Disease*. StatPearls Publishing. January 8, 2024, https://www.ncbi.nlm.nih.gov/books/NBK535419/

McCulloch, W., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*. marlin.life.utsa.edu. https://marlin.life.utsa.edu/mcculloch-and-pitts.html

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nia, N., Kaplanoglu, E., & Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, *3*(1). https://doi.org/10.1007/s44163-023-00049-5

Ochella, S., & Shafiee, M. (2021). Performance metrics for artificial intelligence (AI) algorithms adopted in Prognostics and Health Management (PHM) of Mechanical Systems. *Journal of Physics: Conference Series*, *1828*(1), 012005. https://doi.org/10.1088/1742-6596/1828/1/012005

OpenAI. (2023). GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

Park, S., Bong, J.-W., Park, I., Lee, H., Choi, J., Park, P., Kim, Y., Choi, H.-S., & Kang, S. (2022). Conbert: A concatenation of bidirectional transformers for standardization of operative reports from Electronic Medical Records. *Applied Sciences*, *12*(21), 11250. https://doi.org/10.3390/app122111250

Piccinini, G. (2004). The first computational theory of mind and Brain: A close look at mcculloch and Pitts's "logical calculus of ideas immanent in nervous activity." *Synthese*, *141*(2), 175–215. https://doi.org/10.1023/b:synt.0000043018.52445.3e

Piccinini, G. (2004). The first computational theory of mind and Brain: A close look at mcculloch and Pitts's "logical calculus of ideas immanent in nervous activity." *Synthese*, *141*(2), 175–215. https://doi.org/10.1023/b:synt.0000043018.52445.3e

Pypi.org. (2024). *Contractions*. PyPI. https://pypi.org/project/contractions/

Qiu, Y., Wang, W., Wu, C., & Zhang, Z. (2022). A risk factor attention-based model for cardiovascular disease prediction. *BMC Bioinformatics*, *23*(S8). https://doi.org/10.1186/s12859-022-04963-w

Raleigh , V., Jefferies, D., & Wellings, D. (2022, November 11). *Cardiovascular disease in England*. The King's Fund. https://www.kingsfund.org.uk/publications/cardiovascular-disease-england

Ramsay, S. E., Arianayagam, D. S., Whincup, P. H., Lennon, L. T., Cryer, J., Papacosta, A. O., Iliffe, S., & Wannamethee, S. G. (2014). Cardiovascular risk profile and frailty in a population-based study of older British men. *Heart*, *101*(8), 616–622. https://doi.org/10.1136/heartjnl-2014-306472

Reading Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., & Slotwiner, D. (2021). Systematic review of current natural language processing methods and applications in Cardiology. *Heart*, *108*(12), 909–916. https://doi.org/10.1136/heartjnl-2021-319769

Rezaianzadeh, A., Moftakhar, L., Seif, M., Johari, M. G., Hosseini, S. V., & Dehghani, S. S. (2023). Incidence and risk factors of cardiovascular disease among population aged 40–70 years: A population-based cohort study in the south of Iran. *Tropical Medicine and Health*, *51*(1). https://doi.org/10.1186/s41182-023-00527-7

Rikard, S. M., Kim, B., Michel, J. D., Peirce, S. M., Barnes, L. E., & Williams, M. D. (2022). Identifying individual social risk factors using unstructured data in electronic health records and their relationship with adverse clinical outcomes. *SSM - Population Health*, *19*, 101210. https://doi.org/10.1016/j.ssmph.2022.101210

Roberts, K., Shooshan, S. E., Rodriguez, L., Abhyankar, S., Kilicoglu, H., & Demner-Fushman, D. (2015). The role of fine-grained annotations in supervised recognition of risk factors for heart disease from ehrs. *Journal of Biomedical Informatics*, *58*. https://doi.org/10.1016/j.jbi.2015.06.010

Roman, W. P., Martin, H. D., & Sauli, E. (2019b). Assessment of risk factors for cardiovascular diseases among patients attending cardiac clinic at a referral hospital in Tanzania. *Journal of Xiangya Medicine*, *4*, 18–18. https://doi.org/10.21037/jxym.2019.03.05

Ruan, Y., Guo, Y., Zheng, Y., Huang, Z., Sun, S., Kowal, P., Shi, Y., & Wu, F. (2018). Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: Results from sage wave 1. *BMC Public Health*, *18*(1). https://doi.org/10.1186/s12889-018-5653-9

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and Research Directions. *SN Computer Science*, *2*(6). https://doi.org/10.1007/s42979-021-00815-1

Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Alois Ettlin, D., Grübner, O., Rinaldi, F., & von Wyl, V. (2023). Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digital Health*, *2*(10). https://doi.org/10.1371/journal.pdig.0000347

Singh, P., Haimovich, J., Reeder, C., Khurshid, S., Lau, E. S., Cunningham, J. W., Philippakis, A., Anderson, C. D., Ho, J. E., Lubitz, S. A., & Batra, P. (2022). One clinician is all you need–cardiac magnetic resonance imaging measurement extraction: Deep Learning Algorithm Development. *JMIR Medical Informatics*, *10*(9). https://doi.org/10.2196/38178

Strodthoff, N., Lopez Alcaraz, J. M., & Haverkamp, W. (2024). Prospects for AI-enhanced ECG as a unified screening tool for cardiac and non-cardiac conditions – an explorative study in emergency care. *European Heart Journal - Digital Health*. https://doi.org/10.1093/ehjdh/ztae039

Tian, T., Jin, G., Yu, C., Lv, J., Guo, Y., Bian, Z., Yang, L., Chen, Y., Shen, H., Chen, Z., Hu, Z., & Li, L. (2017). Family history and stroke risk in China: Evidence from a large cohort study. *Journal of Stroke*, *19*(2), 188–195. https://doi.org/10.5853/jos.2016.01270

Turchin, A., Masharsky, S., & Zitnik, M. (2023). Comparison of Bert implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, *36*, 101139. https://doi.org/10.1016/j.imu.2022.101139

Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., & Slotwiner, D. (2022). Systematic review of current natural language processing methods and applications in Cardiology. *Heart*, *108*(12), 909–916. https://doi.org/10.1136/heartjnl-2021-319769

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (n.d.). Attention Is All You Need. *31st Conference on Neural Information Processing System*. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Viani, N., Miller, T. A., Napolitano, C., Priori, S. G., Savova, G. K., Bellazzi, R., & Sacchi, L. (2019). Supervised methods to extract clinical events from cardiology reports in Italian. *Journal of Biomedical Informatics*, *95*, 103219. https://doi.org/10.1016/j.jbi.2019.103219

WHO. (2019). World Health Organization. https://icd.who.int/browse10/2019/en

WHO. (2021). *Cardiovascular diseases (cvds)*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Wu, Y., Fan, Z., Chen, Y., Ni, J., Liu, J., Han, J., Ren, L., Tu, J., Ning, X., & Wang, J. (2019). Determinants of developing stroke among low-income, rural residents: A 27-year population-based, prospective cohort study in Northern China. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00057

Wythoff, B. J. (1993). Backpropagation Neural Networks. *Chemometrics and Intelligent Laboratory Systems*, *18*(2), 115–155. https://doi.org/10.1016/0169-7439(93)80052-j

Yusuf, S., Joseph, P., Rangarajan, S., Islam, S., Mente, A., Hystad, P., Brauer, M., Kutty, V. R., Gupta, R., Wielgosz, A., AlHabib, K. F., Dans, A., Lopez-Jaramillo, P., Avezum, A., Lanas, F., Oguz, A., Kruger, I. M., Diaz, R., Yusoff, K., … Dagenais, G. (2020). Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (pure): A prospective cohort study. *The Lancet*, *395*(10226), 795–808. https://doi.org/10.1016/s0140-6736(19)32008-2

Zhan, X., Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2021). Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns*, *2*(7), 100289. https://doi.org/10.1016/j.patter.2021.100289

Zhang, W., & Cao, T. (2023). Automated Type 2 Diabetes Case and Control Identification from the MIMIC-IV Database. *AMIA Jt Summits Transl Sci Proc.*, 602–611. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10283086/

APPENDIX A LITERATURE SEARCH

Databases	Search Terms	Search Date	Dates	Number of articles retrieved
PubMed	((Natural Language Processing) OR (NLP)) AND ((Cardiology) OR (Cardiovascular Disease) OR (Heart Failure) OR (Myocardial Infarction)) AND ((Clinical Notes) OR (Electronic Health Records) OR (Health Records) OR (Unstructured Notes))	02/04/2024	01/01/2017 – 31/12/2023	195
IEEE Xplore	((natural language processing) OR (nlp)) AND (cardio*) AND ((Clinical Notes) OR (Electronic Health Records) OR (Health Records) OR (Unstructured Notes))	02/04/2024	01/01/2017 – 31/12/2023	18
ACM Digital Library	[[Abstract: natural language processing] OR [Abstract: nlp]] AND [[Abstract: cardiology] OR [Abstract: cardiovascular disease] OR [Abstract: cardio*]] AND [[Abstract: natural language processing] OR [Abstract: nlp]] AND [[Abstract: clinical notes] OR [Abstract: unstructured notes] OR [Abstract: electronic health records] OR [Abstract: ehr]] AND [E-Publication Date: (01/01/2017 TO 31/12/2023)]	02/04/2024	01/01/2017 – 31/12/2023	514

APPENDIX B REGEX FOR DATA CLEANING

Regex	Objectives
`re.sub(r'-?\byears? ?-?old\b|\by(?:o|r)*[ ./-]*o(?:ld)?\b', ' yo', text, flags=re.IGNORECASE)`	*change* 'year old', 'yearsold', etc. *to* 'yo'
`re.sub(r'\b(gentlman|male|man|m|M)(?!\S)\b', 'male', text)`	*Change* 'gentleman', 'male', 'man', 'm', 'M' *to* 'male'.
`re.sub(r'\byr[\'s]*\b', 'years', text)`	*change yr, yrs, yr's to years*
`re.sub(r'\b[P|p]t.?|\b(IN|OU?T) PT\b', 'patient ', text)`	*change Pt and pt with patient, and IN/OUT/OT PT with patient*
`re.sub(r'\b(female|woman|f|F)(?!\S)\b', 'female', text)`	*Change* '`female`', 'woman', 'f', 'F' *to* 'female'.
`re.sub(r'(={2,})', ' ', text)`	*remove "==" from the text (mostly denotes end of the section in the document) that is at least len of 2. So, '=' will not be removed.*

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa