ClinicalICDBERT: Predicting Re-Admission Risk from Clinical Notes, Vital Signs and ICD Codes using BERT Models

by

Merzia Naeem Adamjee

A MS Thesis

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science

APPROVED:

_____

Professor Emmanuel O. Agu, Thesis Advisor

_____

Professor Rodica Neamtu, Thesis Reader

_____

Professor Craig E. Wills, Head of Department

ABSTRACT

Readmissions are a financial burden and challenge for hospitals. Prior work has explored various structured predictors and machine learning algorithms to predict the risk of readmissions due to complications following colorectal, cardiac, and abdominal surgeries [1] and heart failure [2]. Models trained on clinical notes have generally resulted in a much better predictive performance for the hospital readmission task. The goal of this project is to analyze the results achieved in prior work and build a model which predicts readmission risk from a combination of clinical notes, vital signs and ICD codes. This will be achieved by concatenating clinical BERT embeddings created via pre-training on clinical notes, the vital signs data and the ICD codes embedding for each patient's visit to predict readmission within the 30-day time period after discharge. In addition, we will also explore various BERT models including ClinicalBERT that has been pre-trained on discharge summaries and clinical text for the task of predicting readmission within 30 days using the MIMIC-III dataset and rigorously evaluate alternative approaches.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my Thesis advisor Professor Emmanuel O. Agu, for his continuous support towards my Master's study and research, for his patience, motivation, enthusiasm and immense efforts in ensuring my success. His constant feedback helped me shape my research goals and improve my research writing abilities.

I would like to extend my appreciation to Professor Rodica Neamtu for being my thesis reader and for her valuable and insightful comments and input on my thesis.

I would also like to thank and acknowledge the efforts of my lab mate Atifa Sarwar, PhD candidate, working under the supervision of Professor Emmanuel O. Agu, for her continuous support, encouragement and guidance throughout the course of my research. The stimulating discussions and knowledge, helped steer my research in the right direction.

Last but not least, I would like to thank my family: my parents Naeem Adamjee and Fatima Adamjee, and my brother, Azeem Adamjee, for being my constant emotional and spiritual support.

NOMENCLATURE

BERT	Bidirectional Encoder Representations from Transformers
ICD	International Classification of Disease
EHR	Electronic Health Records
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
TL	Transfer Learning
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
NLP	Natural Language Processing
ML	Machine Learning
HHC	Home Healthcare
ITL	Inductive Transfer Learning
AUROC	Area Under the Receiver Operating Characteristic curve
AUPRC	Area Under the Precision-Recall curve
RP80	Recall-Precision at 80%

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Introduction

Hospital readmission is among the most critical issues in the healthcare system due to its high frequency and cost. A hospital readmission is defined as an admission to a hospital shortly after discharge within a short period of time (typically within 30 days) after an original admission [3] where the patient requires retreatment. Readmitted patients represent a significant health care value problem and are associated with high financial costs, which ultimately make the high readmission rates undesirable. Readmissions can be for planned or unplanned reasons [3] and could be at the same hospital as original admission or a different one. Since readmissions are an economic burden on the healthcare system, it is necessary that hospitals reduce their readmission rates in a cost effective manner while identifying patients who indeed require readmission after being discharged. Approximately 10% of critically ill patients following discharge face the adverse outcome experience of being readmitted [4]. Such a high rate of readmission is also an indicator of poor or incomplete medical care [4] provided by the hospital.

Electronic Health Records (EHRs) contain huge amounts of patient information including patient physiological data, interventions, treatments and diagnoses [5]. EHRs also include clinical notes, vital sign data and ICD code information. Clinical notes written by clinicians and caregivers provide a much richer insight of a patient's symptoms, radiology results, daily activities and patient history since the time of admission [6]. In addition, the instability of vital signs during a patient's stay and at the time of discharge can also be used as an important feature to assess readiness and safety of discharge, and to lower 30 day mortality and readmissions [7]. Furthermore, ICD (International Classification of Disease) coding, which are alphanumeric codes assigned to every disease, injury and symptom, are used by doctors, healthcare consulting firms and public health agencies all over the globe to represent different diagnoses. They are considered one of the most predictive variables that classify and encodes all diagnoses, symptoms and procedures associated

with a patient's visit, which can aid in hospital readmission prediction.

In prior work, methods have been developed to predict the risk of readmissions due to complications following colorectal, cardiac, and abdominal surgeries [1] and heart failure [2]. These previous efforts to predict readmission risk, have relied mostly on structured predictors [1] and machine learning algorithms. Many studies have trained ML classifiers such as Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) [8] to obtain a patient's risk of readmission. Some have even considered this problem as a binary classification task [3], while others have stated the need for a solution tailored for specific disease groups for better prediction. Models trained on clinical notes ([6],[9]) have indeed resulted in a much better predictive outcome for the hospital readmission task. Very few prior approaches have combined structured and unstructured data as inputs to a model, which we explore.

With the increase in free and unstructured clinical text data available nowadays and the predictive ability of ICD codes and vital signs, our main contribution in this work will be to develop a model and evaluate it's predictive capabilities on 30-day readmission using all three data types as input features to a ML model. In this research, we will utilize the MIMIC-III dataset, a large database comprising of information relating to patients admitted to critical care units, to build a BERT (Bidirectional Encoder Representations from Transformers)-based model that can perform predictions for our specific task of patient readmission within a 30 day time period using clinical notes, vital signs and ICD codes. A transformer [10] is a deep learning model that adopts the mechanism of self-attention and assigns a different weight to each part of the input. Similarly, BERT is a transformer-based language representation model designed to pre-train deep bi-directional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT models are good at handling language-based tasks and have outperformed many models on simple NLP tasks. In rigorous evaluation, we will compare our proposed model with previously proposed BERT and Machine Learning (ML) baseline models including ClinicalBERT [6], which are text representations that were developed from clinical notes to leverage the temporal information of the sequence of the notes. Both [6] and [9] use physician and nursing notes from the

MIMIC-III dataset. These models represent the state-of-the-art in readmission risk prediction, and have outperformed the baseline models BERT [11], XLNet [12], highlighting the need for building models tailored to clinical data. To assess the performance differences amongst the models, we will use the standard precision, recall, F-measure evaluation metrics as well as AUROC, AUPRC and RP80.

1.2 Importance of using high-quality clinical notes to make predictions

The use of clinical notes to predict clinical outcomes has become widespread. These notes impart rich information regarding a patient's medical and socio-behavioral risk factors [13]. In [13], each clinical nursing note documented for a Home Healthcare patient is classified as either "concerning" or "not concerning", using a rule-based NLP system that identifies the various risk factors that may lead to a patient's risk of hospitalization. Another study [14] leveraged both longitudinal electronic health records and clinical documentation to improve predictive models of HIV diagnosis. This study showed that NLP models were able to extract clinical terms that were highly indicative of high-risk behaviour. The predictive model which combined the baseline and NLP keyword model, yielded the highest precision, an F-measure of 0.74 much better than the baseline model itself that did not include keywords indicative of additional HIV risk factors [14].

1.3 Importance of Real-time EHR data

The use of electronic health records (EHRs) has sky-rocketed in the past couple of years. In 2009, 12.2% of US hospitals had a basic EHR system, which has increased to 75.5% by 2014 [15]. Many machine learning techniques are applied to EHRs data to identify patients at high risk of serious conditions. The adoption of digital records over paper records is mainly due to how patient data is organized digitally and made accessible. EHRs offer access to longitudinal data that may aid in predicting future outcomes or diagnoses, and open opportunities to personalize decision-making for a given patient [16]. These comprehensive digital records store both structured and unstructured data and its real-time recording has shown to improve patient care greatly.

1.3.1 Incorporating vital signs data for better prediction

Vital signs are measurements of a body's most basic functions. There are four main vital signs that are routinely monitored by clinical professionals and health care specialists which include: body temperature, pulse rate (heart rate), blood pressure and breathing rate. The utility of these vital signs has been majorly neglected and underutilized to identify disease severity in patients. Therefore, in [17], the authors show that an increase in the respiratory rate can have a significant impact on patients who suffer from sepsis and early and timely interventions with the help and incorporation of vital signs data can improve the outcome for the patients.

1.4 Measuring patient health using ICD codes

International Classification of Diseases (ICD) code is an important label of EHR. They are used to identify symptoms and signs, diseases and abnormal findings, or operations as general labels for clinical records [18]. A study conducted in [19], compared the predictive power of the uniformly available ICD codes to Injury Severity Score (ISS) to derive survival risk ratios, and confirmed that ICD diagnosis codes could be equal or better predictors of survival. Another study [20], sought to determine the accuracy of ICD-9-CM(Clinical Modification) codes for identifying cardiovascular and stroke risk factors and concluded that ICD-9-CM codes alone do not guarantee the presence of heart failure, coronary artery disease, diabetes, hypertension and stroke and requires additional data to make solid conclusions. However, the study did conclude that ICD-9-CM codes for all risk factors had high specificity (>0.95) and low sensitivity (<=0.76).

1.5 Impact of structured and unstructured data on prediction

As discussed in section 1.3, EHRs are gaining rapid adoption due to their rich, longitudinal qualities and their versatility to contain both structured and unstructured data. Structured data is easier to extract and incorporate into a predictive model than unstructured data. But recent advancements in NLP, have opened new opportunities to developing techniques that have allowed the easy extraction of valuable information from unstructured text in EHR progress notes [21]. In addition, the authors demonstrated in their study, [21], that both unstructured and structured

information combined, have exhibited interesting patterns that enhance the predictive ability of the model and facilitate early detection of diseases such as heart failure. The idea that NLP is more focused on word, sentence and document-level representations of unstructured text and that patient health data can be represented in structured forms such as medical codes, demographics and vital signs, gives a monumental launching off point for future deep clinical research [22].

1.6 Advantages of models that utilize Transfer Learning

In recent advancements, the field of NLP has witnessed the rise of several transfer learning methods and architectures being applied in the field, which have remarkably improved upon state-of-the-art models on a number of NLP tasks. The idea behind transfer learning, is to transfer parameters or knowledge from one trained model to another [23]. Modern transfer learning methods focus on the pre-training stage, understanding the information that the learnt representations capture and adapting these models to downstream NLP tasks. Usually architecture modifications entail adding a few additional embeddings to additional layers on top of a pre-trained model or inserting intervening layers or modules inside the pre-trained model [24]. TL is divided into transductive transfer learning and inductive transfer learning (ITL). We will be focusing on ITL which is further divided into four categories: Sequential fine tuning, adapter modules, feature-based and zero-shot. The sequential fine tuning approach, is regarded as one of the dominant approaches in the field as compared to the other approaches. BERT's architecture follows the sequential transfer learning approach and is trained on two unsupervised tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP). Thus fine-tuning simply entails providing the model with the required inputs and outputs representations that are compatible for the tasks.

1.7 Thesis overview

In order to measure the performance of clinical notes with ICD codes, we proposed a model that used the embeddings of both input features to predict hospital readmission. We evaluated our model against the state-of-the-art ClinicalBERT model and previously proposed ML classifiers.

We incorporated the learnings from the previously proposed models and utilized the data from

the MIMIC-III dataset in this study. Feature selection is an important task to maximize the target task's performance.

1.8 Novelty of this thesis in relation to prior work

This thesis explores the importance of predicting hospital readmissions using clinical notes, EHRs and ICD codes. Previously no studies have incorporated all three features using a BERT model to predict readmissions. There have been studies that have investigated common causes and pattern of short and long term readmissions in stroke patients [25]. Many other have utilized machine learning algorithms to predict readmission but none used the BERT approach with these input features.

1.9 Thesis contributions

- We proposed a method to combine structured and unstructured data using BERT models
- We proposed a set of embeddings as representations for combining structured and unstructured data types
- Systematic evaluation of our approach against the ClinicalBERT baseline model for predicting hospital readmission including rigorous experiments with ML classification baselines as comparisons to our new model

1.10 Thesis outline

The thesis is presented in subsequent chapters as follows. Chapter 2 summarizes a background of hospital readmission, followed by prior works that have tackled the readmission prediction task as well. In Chapter 3, we proposed a model for readmission prediction that uses ICD codes and clinical notes, the MIMIC-III dataset and evaluation methods. The evaluation results are presented in Chapter 4 with discussion in Chapter 5. Finally, Chapter 6 concludes this study and discusses any possible future work.

2. BACKGROUND AND RELATED WORK

2.1 Hospital Readmission

Multiple factors can prompt a reason for a patient's readmission to a hospital and many factors such as age, co-morbidities, economic disadvantage and number of previous admissions have been used as predictor variables to predict hospital readmissions. But the outcome obtained from these models have not been very conclusive due to the non-homogeneity of the patient groups [26]. Moreover, studies ([27],[28]) have found that the most common reasons for readmission have been heart failure [29], gastrointestinal disorders, pneumonia[30] and acute myocardial infarction (AMI) [31]. These diseases pose a substantial health and economic burden nationally and readmissions due to them are expected to grow significantly in the next two decades in the United States [31]. The use of data obtained from inpatient care records, discharge notes and clinical data have shown to reduce preventable readmissions more than the above mentioned factors.

2.2 Machine learning models used to predict hospital readmission

Recent comparative studies have suggested that machine learning techniques can improve prediction ability over traditional statistical approaches [32]. Performance of ML models used for readmission risk prediction can be improved by incorporating feature selection techniques and class imbalance. Feature selection techniques allow the identification of the most important variables of a dataset and this has proven useful in identifying the key variables associated to a disease. In addition, feature selection techniques aid in reducing overfitting and the complexity of the model, increasing model interpretability. Class imbalance on the other hand deals with an imbalanced training dataset, that causes a resulting model to be biased to a majority class. ML techniques such as Logistic Regression and neural network models have proven to be the more dominant modelling methods [32].

2.3 Prior work predicting hospital readmission

This section reviews previous work on hospital readmission prediction of patients using machine learning classifiers including Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Multi-Layer Perceptron (MLP), ClinicalBERT [6] and ClinicalXLNet [9].

A similar study conducted in [8] explores the MIMIC III dataset, in particular the International Classification of Disease (ICD) codes, vital signs from chart events and gender from demographics data to target the problem of 30-day readmission prediction. Before generating the feature vectors, they identified the patient visits that were readmissions. For each visit they extracted the features from the database (vital signs and ICD codes) and performed aggregation at visit level for both the vital sign measurements and ICD code embeddings, which were used as input to the classifiers. They trained four classifiers, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and Multi-Layer Perceptron (MLP), of which Random Forest gave the best results of accuracy 0.656 and area under the ROC curve of 0.660. The limitations with SVM, LR and MLP were that they had a higher false positive rate compared to RF, with most visits incorrectly classified as readmission.

Another study [3], formalized the task of predicting early patient readmissions as a binary classification task using the New Zealand National Minimum dataset, obtained from the New Zealand Ministry of Health. They used a few background variables such as a patient's race, sex, age, and length of stay, and the most informative aspect of the dataset as noted by them, the large collection of ICD-10-AM (Austria modification) medical codes assigned to each patient visit. Just as the previous study, Random Forest and deep neural networks have significantly better predictive performance than Logistic Regression. While deep learning models pose the most challenges due to the large number of model parameters, they also possess the greatest potential to boost predictive accuracy in statistical approaches to predicting early admission [3].

While [3] is a study that was done in 2015, they identified that "local" methods that were specific to a particular patient population outperformed a "global" method that did not take in to account the varying nature of the different disease groups. The benefit of this as noted by them is

that the model would be tailored to the specific population of interest and consequently result in a better prediction of readmissions of specific subgroups.

Barbieri et al. [4], in their study compared several deep learning models for predicting the risk of readmission within 30 days of discharge from the ICU, based on the full clinical history of a patient. They concluded from their research that attention-based networks may be preferable to recurrent networks and that diagnosis and procedure codes associated with ICU admissions prior to the index admission, as well as medications and vital signs, alone provide limited additional value when predicting readmission.

The more recent work [33] on readmission prediction has used clinical notes to predict readmission within 48 hours of a patient being admitted to the hospital. The proposed predictive model showed better prognostic capability in predicting readmission than the machine learning model alone. They showed significant improvements of 11%-28% in sensitivity and 1%-3% in the area-under-the-receiver operating characteristic (AUROC) curve. Also, as mentioned in [34], prediction models using EMR (Electronic Medical Records) data have better predictive performance than those using administrative data.

ClinicalBERT [6] and ClinicalXLNet [9] made use of clinical notes which is unexploited in predictive modeling compared to structured data. In [9] the clinical notes leveraged temporal information of the sequence of the notes and their experiments demonstrated that Clinical XLNet outperformed the best baselines consistently.

Our work will differ from all prior work because we are focusing on combining vital signs with embeddings generated from two types of input data (clinical text and ICD codes) to predict readmission and to investigate whether the aforementioned information will produce better readmission risk prediction using the BERT architecture and compare our results with the previously pre-trained BERT ([6]) model.

2.4 Background

2.4.1 Deep Learning and Natural Language Processing (NLP)

Natural Language Processing (NLP) with Deep Learning (DL) is a powerful combination. With the creation of neural language models the process of obtaining good word embeddings has become much easier. The first word embedding model utilizing neural networks was published in 2013 [35] by research at Google. Word embeddings have become an integral part of NLP models nowadays, mainly because of their effectiveness in modeling the semantic importance of a word by placing semantically similar inputs close together in the embedding space and in a numeric form, which makes it easier to perform mathematical operations [36]. Embeddings allow tranforming high-dimensional vectors into a relatively low-dimensional vector.

Word embeddings which is a learned representation of text where individual words are represented as real-valued vectors in a predefined vector space and words that have the same meaning have a similar representation. This has been one of the key breakthroughs when it comes to the outstanding performance of deep learning methods on natural language processing tasks. In addition, the word representations computed using neural networks as stated in [35], is particularly exciting because the learned vectors explicitly encodes many linguistic regularities and patterns. Simple word embedding approaches have performed quite well, but the selection of a word embedding technique is based upon task-specific requirements and careful, thorough experimentation, as well as fine-tuning the word embedding models to improve overall accuracy.

In our work we will be generating embeddings for the clinical text data as well as the ICD codes. We will use embeddings as part of our project because we intend to capture the meaningful information drawn from the context of the clinical notes and this approach has proven to make textual data understandable to machine learning and NLP models. In addition, embeddings make it easier to do machine learning on large inputs and can be learned and reused across models.

2.4.2 BERT:Overview

Bidirectional Encoder Representations from Transformer (BERT) is a language representation model designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The BERT framework can be divided into two steps: pre-training and fine-tuning. During the pre-training phase the model is trained on unlabeled data over different pretraining tasks. While during the fine-tuning step, the BERT model is first initialized with the pre-trained parameters, and subsequently all of the parameters are fine-tuned using labeled data from the downstream task.[11]

BERT eliminates the major shortcoming of the previous standard language models that are unidirectional by using a ”Masked Language Model” (MLM) pre-training objective. The MLM procedure allows some percentage of the input tokens to be masked at random, while consequently predicting these same masked tokens. Too little masking might be too expensive to train while too much masking might not have enough context. 80% of the time the masked token is replaced with [MASK], 10% of the time with a random word and the remaining 10% with the same word. In addition, the BERT model also utilizes the ”Next Sentence Prediction” task that jointly pre-trains text-pair representations to understand the relationship between two sentences. When choosing sentences A and B, half of the time sentence B is the actual sentence that follows sentence A, while the other 50% of the time it is a random sentence from the corpus. Figure 1 gives a visual overview of BERT’s framework on downstream tasks. [11]

Each input token is represented as a sum of three embeddings: token embeddings, segment embeddings and position embeddings. BERT has an embedding layer that enables the conversion of each word into a fixed length vector of defined size. The attention mask indicates which inputs require additional attention and which do not. Self-Attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. This inturn helps in avoiding locality bias which means that long distance context has "equal opportunity".

Consequently, the encoder layer reads the text input and produces a prediction for the task. The

"bert-base" architecture is composed of 12 successive transformer layers, each having 12 attention heads and 768 hidden features. [11]. A distinctive feature of BERT is its unified architecture across different tasks as well as its ability to produce word representations that are dynamically informed by surrounding words, regardless of context. This new state-of-the-art model is conceptually simple and empirically powerful.

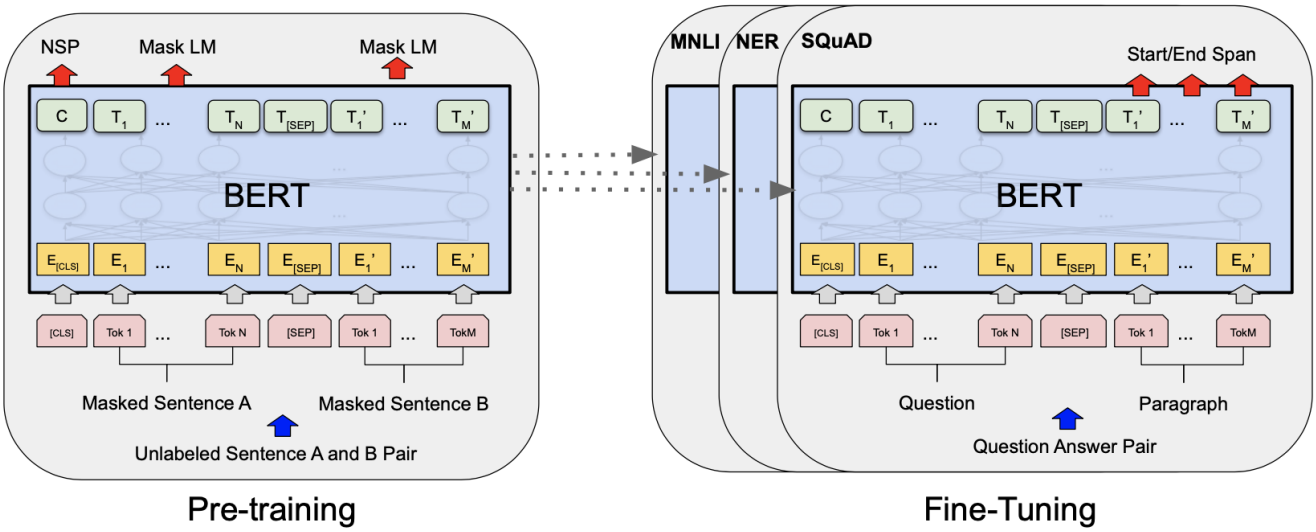

Figure 2.1: Overall pre-training and fine-tuning procedures for BERT. The same architecture is used for pre-training and fine-tuning. [11]

2.4.3 Need for domain-specific BERT models

The BERT framework has been pre-trained using text from Wikipedia. Applying the same model to text that is domain specific (e.g clinical, legal, financial) will not give a high accuracy result and will require continuing training BERT with some of the text data or on a more domain-specific language model. Some domain-specific language models that have been created by training the BERT architecture from scratch on a domain-specific corpus that generate embeddings better suited for the domain-specific NLP problems, include:

BioBERT (biomedical literature corpus): Domain-specific language representation model pre-

trained on a large scale biomedical corpora [37]

ClinicalBERT (clinical notes corpus): Pre-trained BERT using clinical text from approximately 2M notes in the MIMIC-III database and fine-tuned on the task of hospital readmission [6]

mBERT (corpora from multiple languages): A Multilingual BERT (mBERT) which provides sentence representations for 104 languages, making it useful for many multi-lingual tasks. [38]

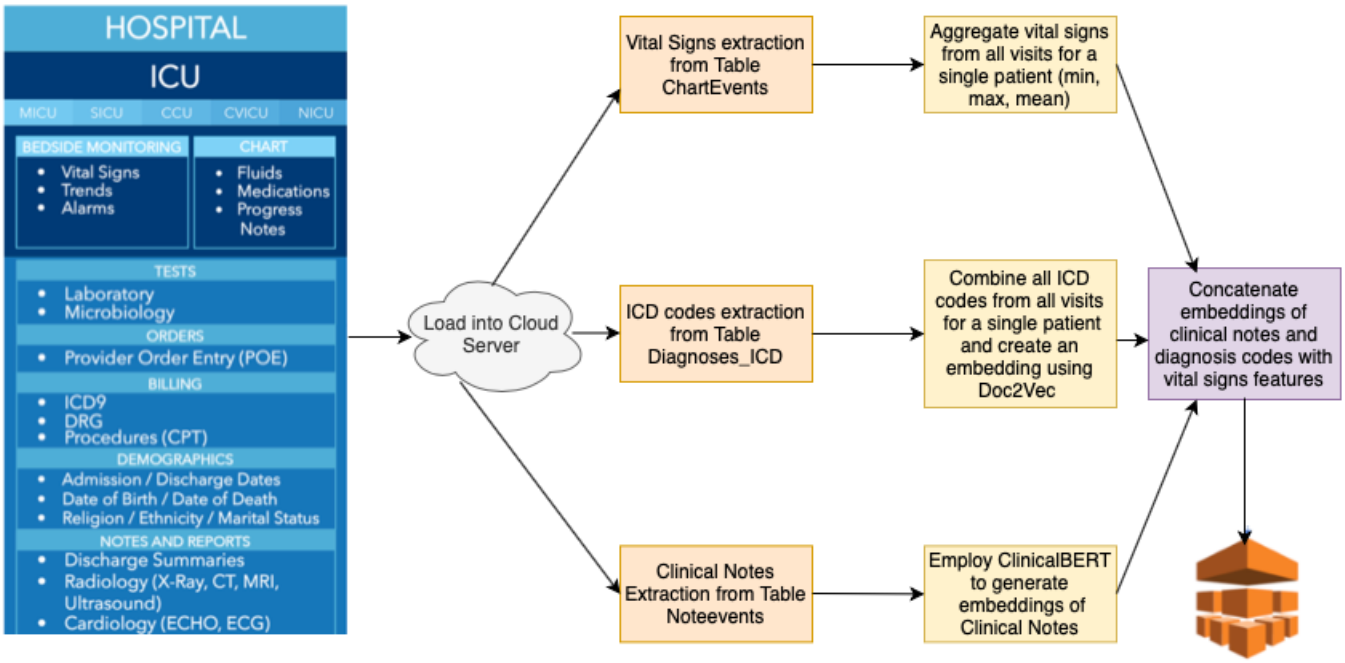

Figure 3.1: This diagram highlights the steps for preprocessing and combining the structured and unstructured data used to train our ClinicalICDBERT model

3.1 Dataset

The Medical Information Mart for Intensive Care (MIMIC-III) dataset is a publicly available critical care database maintained by the Massachusetts Institute of Technology (MIT)'s Laboratory for Computational Physiology [39]. The availability of the MIMIC-III dataset has enabled researchers to benchmark machine learning models for studying clinical outcomes such as length of hospital stay [6] and mortality. It is a relational database consisting of 26 tables comprised of deidentified, comprehensive clinical records of patients admitted to an Intensive Care Unit (ICU) during 2001 and 2012 at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Mas-

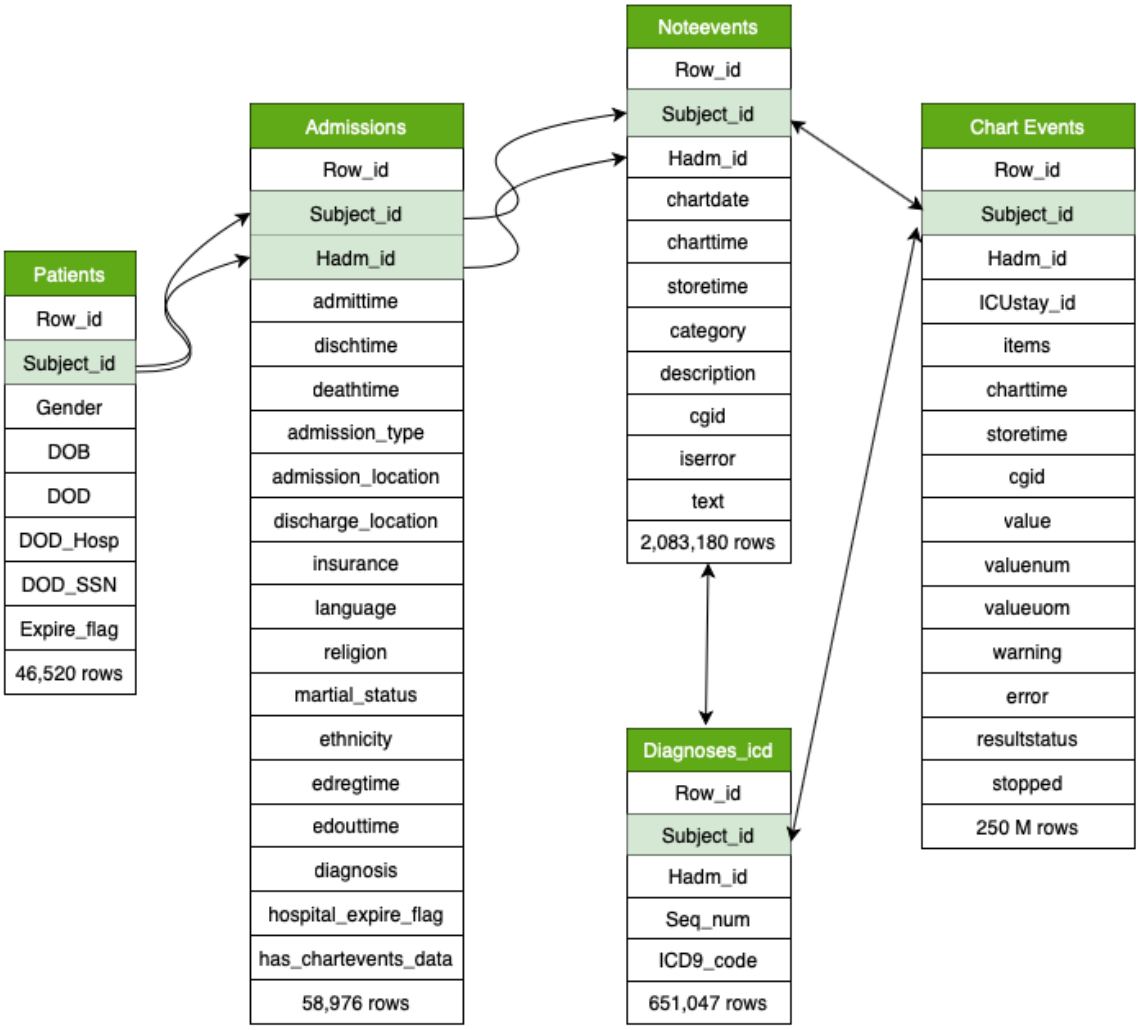

Figure 3.2: The schema of tables from the MIMIC-III dataset

sachusetts. The database includes information relating to demographics, vital sign measurements made at bedside (approximately 1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging notes and mortality. MIMIC-III contains data associated with 53,423 distinct hospital admissions for adult patients, aged 15 years or above and 7,870 newborns admitted to an ICU at BIDMC. In addition, it has 2,083,180 recorded clinical note events [9].

3.2 Research Methodology

For this thesis, the research focused on developing a model for learning deep representations of a combination of clinical text, vital signs and ICD codes as seen in Figure 3.3. We employed a pre-trained ClinicalBERT that was initially trained on the BERT model using all the clinical notes from MIMIC III. These notes can be found in the NOTEEVENTS table in the database. We also used the vital signs and ICD codes data from the MIMIC-III database which can be found in the CHARTEVENTS and OUTPUTEVENTS, and DIAGNOSES ICD and PROCEDURES ICD tables respectively. We understood the embedding layers, attention mask and encoder layer of the ClinicalBERT model to develop a model that further explores the impact of all three of our inputs. A token from the clinical text data will be represented as a sum of the token embedding, a learned segment embedding and a position embedding. The embedding layer preserved the different relationships between the text in the clinical data including semantic, syntactic, linear and contextual relationships. Similarly, ICD codes were mapped to "embeddings" using the Doc2Vec embedding technique. And consequently, concatenate vital signs data with the embeddings of the clinical notes and ICD codes.

These embeddings will be further processed by the attention mechanism, the dot-product attention, to compute a weighted average of the text and code embeddings separately, where a higher weight is assigned to the most relevant text and codes. Lastly, the computed clinical embeddings, diagnoses/procedures and vital signs will be concatenated to a vector of static variables and passed to a fully connected layer with a sigmoid activation layer to the output of the BERT model. The output of the model will correspond to the risk of readmission to the hospital within 30 days of discharge.

3.3 Baseline Models

We conducted a set of experiments against the state-of-the-art ClinicalBERT baseline and ML classifiers:

- ClinicalBERT is a modified BERT model, pre-trained on MIMIC-III patient clinical notes,

used for predictive downstream tasks. The model outperforms BERT by uncovering high-quality relationships between medical concepts and jargon

- Random Forest is a supervised machine learning algorithm that is used widely in classification
- Gradient Boosted Machines (GBM) is a machine learning technique used in regression and classification tasks
- Ada Boosting is a popular boosting technique that aims at combining multiple weak classifiers to create one strong classifier
- SVMs are a set of supervised learning methods used for classification, regression and outliers detection
- Naive Bayes is from the family of probabilistic classifiers which is used for the classification task
- k-Nearest Neighbor (kNN) is one of the simplest, yet efficient supervised learning algorithms used to solve classification

3.4 Evaluation Metrics

To measure the performance of the models in predicting 30-day readmission, we used the area under the receiver operating characteristic (AUROC) curve, the precision-recall curve, accuracy and RP80 (Recall Precision at 80%) at specific thresholds. These are all calculated using false positive (FP), true positive (TP), true negative (TN) and false negative (FN) as follows:

- Area under the receiver operating characteristic curve (AUROC): The area under the true positive rate versus the false positive rate
- Area under the precision-recall curve (AUPRC): The area under the plot of precision versus recall

- Recall at precision at 80% (RP80): Fixing precision at 80% (or, 20% false positive in the positive class predictions). This threshold is used to calculate recall. This metric enables building models that minimizes the false positive rate.

- Accuracy: It is calculated as the ratio between the number of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{3.1}$$

- Recall: The measure of the model correctly identifying true positives

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3.2}$$

- Precision: Number of true positives divided by the total number of true positives plus the number of false positives

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3.3}$$

3.4.1 Evaluation Method

Holdout validation: Hold-out validation includes splitting the dataset into a 'train' and 'test' set. The training set is used to train the model and the test set is used to evaluate the model on unseen data. The data was split into train, validation and test sets of respectively 80%, 10% and 10% splits. The main idea of splitting the dataset into a validation set is to prevent our model from overfitting i.e., the model classifies samples in the training set accurately but does not generalize over data it has not seen before and mis-classifies.

3.5 Embeddings

3.5.1 ICD-9 codes embeddings

In the MIMIC-III dataset each patient has a unique "Subject_ID". Each hospital admission of a patient has a separate admisssion ID denoted by "HADM_ID" (Hospital Admission ID). Which means one subject_id can be associated with multiple hadm_ids when a patient has had multiple admissions. The "DIAGNOSES_ICD" table contains tuples of patient ID (subject_id), admission ID (hadm_id), and ICD-9 code. We created a list of all the ICD codes linked with a single unique patient ID and used Doc2Vec to generate a representation of the ICD codes for each patient. Doc2Vec is an unsupervised method and the goal of doc2vec is to create a numeric representation of a document, regardless of its length. But unlike words, documents do not have a logical structure. Therefore, the concept suggested by Mikilov and Le [40] was that they used the word2vec model, and added another feature vector, which is document-unique. So in case of ICD codes, when we trained the list of codes, at the end of training, we obtained a numeric representation for all the ICD codes linked with each patient.

3.5.2 Pre-trained ClinicalBERT embeddings

ClinicalBERT has been pre-trained on clinial notes. For our experiments we utilized the pre-trained ClinicalBERT embeddings directly instead of training the model again from scratch.

3.6 Vital Signs

While we have not yet included vital signs in our experiments, we plan to incorporate it as part of future work as part of this thesis.

3.7 Experimental Setup

We evaluated the proposed ClinicalICDBERT on the readmission prediction task using AUROC, AUPRC and RP80. We compared two types of models: (1) existing state-of-the-art ClinicalBERT model pre-trained on the MIMIC-III clinical notes and fine-tuned directly on the readmission task; and (2) refined ClinicalICDBERT fine-tuned using ICD codes on the readmission

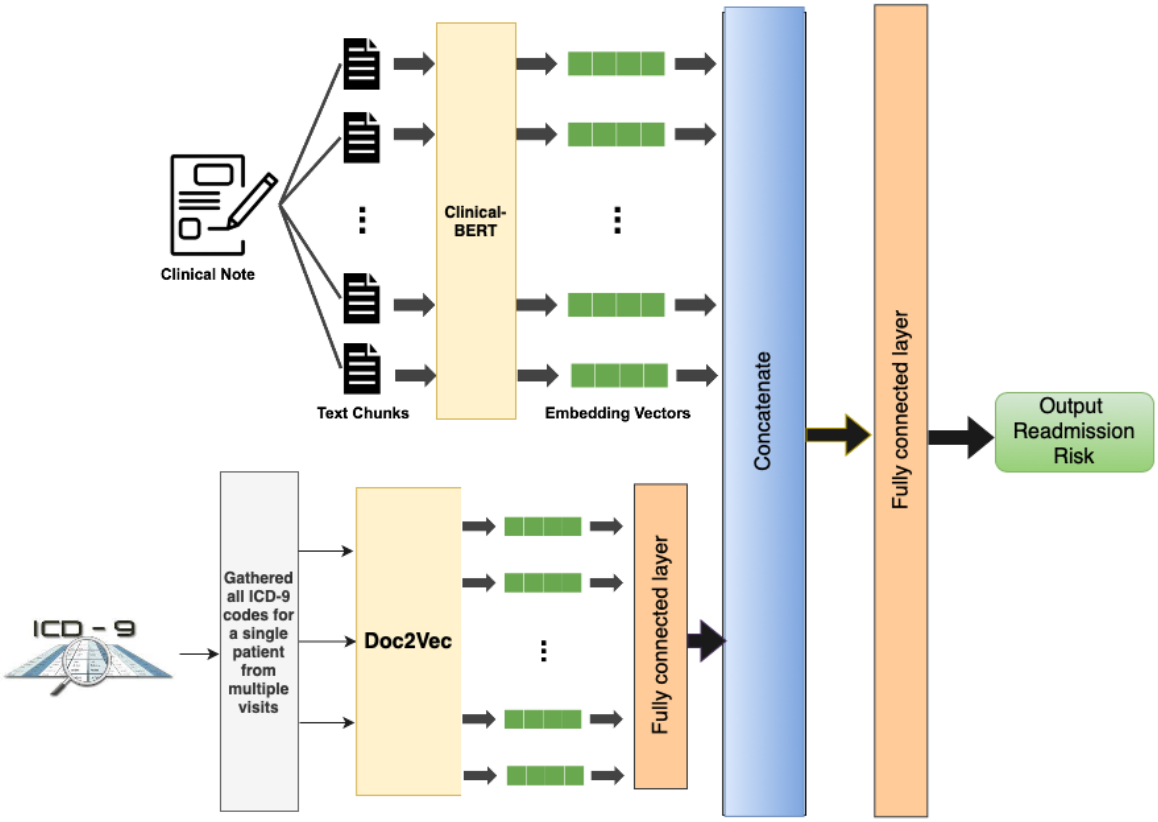

Figure 3.3: This diagram shows the overview of our proposed approach

task.

3.7.1 Datasets

We evaluated the performance of the ClinicalICDBERT model on three MIMIC-III dataset features: (a) clinical notes (b) ICD codes (c) vital signs. Table summarizes the MIMIC-III dataset. For a fair comparison between both models, we used the same training, development and test sets to train and evaluate the models. MIMIC-III consists of 2,083,180 rows of de-identified clinical free-text notes. In the DIAGNOSES_ICD table there are 172335 unique admission IDs.

We experimented with the readmission prediction based on discharge notes.

- Discharge summary notes: This clinical note set contains discharge summaries. A discharge summary is a written summary of a patient's status during an admission. It includes valu-

able information such as reasons of hospitalization, diagnosis, a complete list of prescribed medications, and a radiology report [41]. The discharge summaries set that we will be utilizing for our experiments contains 6162 unique admissions. We split the discharge summary associated with a patient into text chunks of size 512.

3.7.2 Baseline

As our baseline model, we use the ClinicalBERT and perform supervised fine-tuning of its parameters for the readmission task. ClinicalBERT serves as a good baseline as its one of the inital models to be trained on clinical notes and provides a single LM to adapt to different domain features. In addition, we also considered ML classifiers including Gradient Boosting, Ada Boosting, SVM, Naive Bayes, Random Forest and kNN as additional baselines and compared our results with them which fared much better than the ML classifiers.

3.7.3 Training

Our implementation of ClinicalICDBERT is based on the work of [6]. We employed ClinicalBERT to generate embeddings for the clinical notes which used the AdamW, a variant of the Adam optimizer with weight decay as the optimizer with a learning rate of $2e^{-5}$ and a batch size of 32. In addition, all the tokenized texts were chunked to span no longer than 512 tokens and the model regularized with dropout of 0.1. In similar fashion, we fine-tuned our model, ClinicalICDBERT, on the discharge summary notes in the MIMIC dataset and the corresponding ICD codes of each patient. We then passed the clinical notes embeddings and ICD codes embeddings through a dropout layer and concatenated both outputs to predict readmission.

3.7.4 Classification Architecture

Following standard practice we passed the final layer [CLS] token representation to a task-specific feed-forward layer for prediction.

3.7.5 Fine-tuning

For fine-tuning, different model hyperparameters were experimented with to observe changes across both BERT variants, ClinicalICDBERT and ClinicalBERT. For our hospital readmission task, we used a learning rate of $2e^{-5}$ and a batch size of 8. We trained the results for 10 epochs in total and performed early stopping, choosing the best checkpoint based on validation set performance (evaluating every 3200 optimization steps). And we used 10% of the training set as validation split.dWe then compared the performance of the ClinicalBERT model with our model. ClinicalICDBERT was trained using the PyTorch library for machine learning in Python. Training our model for 10 epochs took 9 hours on the three tables from the dataset.

3.7.6 Hyper-Parameter Tuning

Table 3.1: Hyperparameters for clinicalICDBERT hospital readmission classifier

Hyperparameter	Assignment
Number of epochs	3
Patience	2
Batch size	8
Learning Rate	$2e^{-5}$
Dropout	0.1
Feedforward layer	1
Classification Layer	1

4. RESULT

One of the key aspects of building practical models is it's fast adaptation to new domains. The purpose of our research was to analyze and explore the results presented in the ClinicalBERT paper and test our own model ClinicalICDBERT. To achieve this we decided to utilize clinical notes and ICD codes to develop a model and evaluate it's predictive capabilities on 30-day readmission using these input features. All three features were extracted using the MIMIC-III dataset tables and were linked using the "subject_id" of each patient. Table 4.1 shows the results of each model on the readmission task.

Table 4.1: Results

Model	AUROC	AUPRC	RP80	Precision	Recall	F-1 Score
ClinicalBERT	0.851	0.83	0.662	0.738	**0.748**	0.743
ClinicalICDBERT	0.851	0.83	**0.665**	**0.740**	0.746	0.743
Gradient Boosting	0.492	0.498	-	0.529	0.124	0.51
Ada Boosting	0.490	0.489	-	0.484	0.417	0.49
SVM	0.540	0.512	-	0.25	0.5	0.5
NB	0.492	0.500	-	0.486	0.444	0.49
KNN	0.525	0.54	-	0.512	0.459	0.52
Random Forest	0.492	0.494	-	0.25	0.5	0.5
Assaf et al.[8]	0.66	-	-	-	-	-

Readmission prediction is a hard problem to solve and much work has been previously done to improve results using BERT models and ML classifiers. The results presented in this section are similar to the ones achieved by ClinicalBERT and can be improved with a little more work. Our model's area under the ROC curve came about to be 0.851, the Precision-Recall curve is 0.83 and the RP80 0.665. We added a dropout layer of 0.1 on the clinical notes and the ICD codes and passed the ICD codes through a linear fully connected layer before making the final readmission prediction by another fully connected layer with linear activation for binary classification.

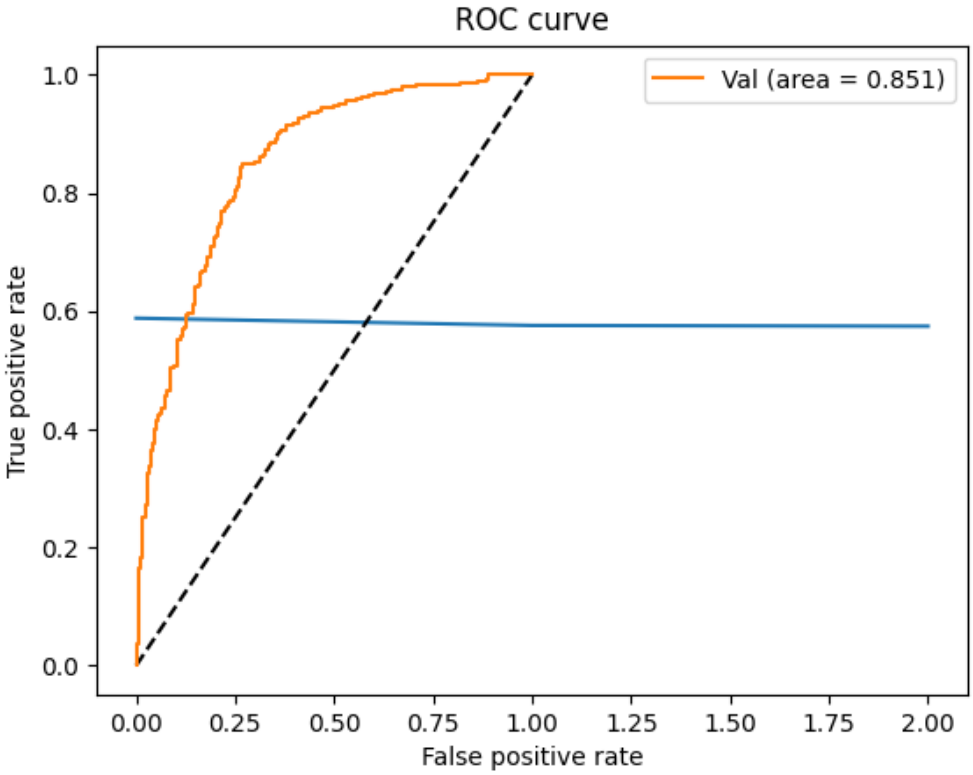

Figure 4.1: The area under the ROC Curve for the ClinicalICDBERT

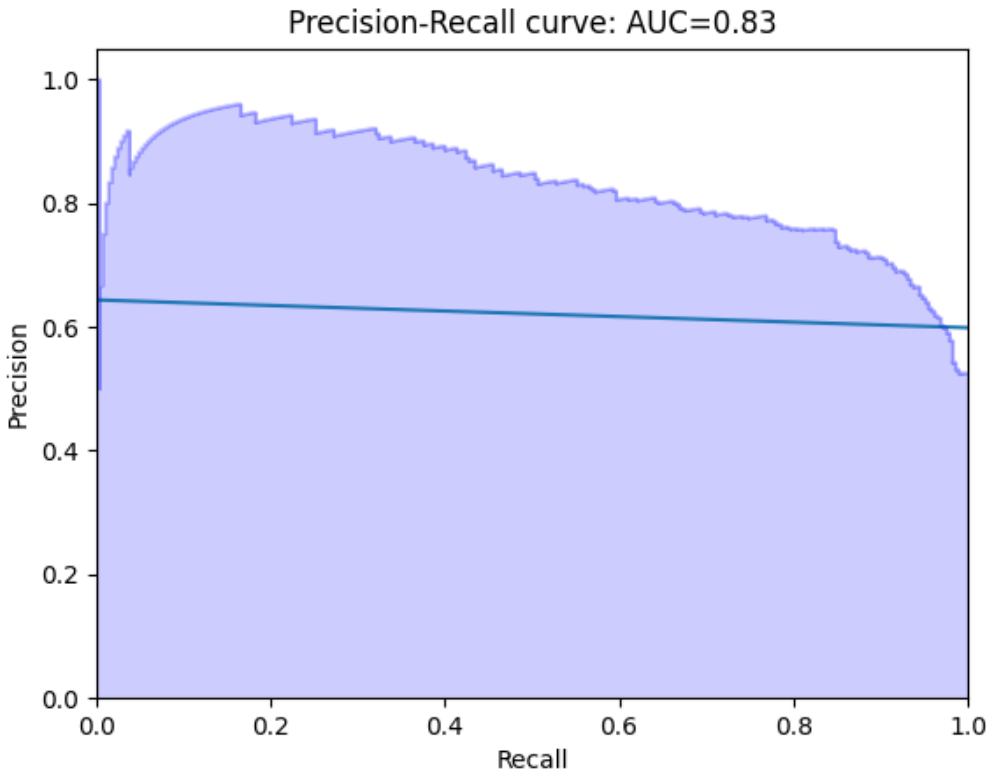

Figure 4.2: The Recall-Precision Curve for the ClinicalICDBERT

5. DISCUSSION

According to widely available research, our thesis attempts to explore and produce results on the readmission prediction problem based on MIMIC. There are still lots of limitations and improvements that can boost the model performance. A step towards better prediction could include improving the type of clinical notes used to create embeddings. Since we only utilized discharge summary notes, using early admission notes in addition to discharge notes could improve the clinical embeddings generated using ClinicalBERT.

Although the use of ICD codes with clinical notes in the prediction of hospital readmission did not significantly surpass the results achieved by ClinicalBERT but we did see an improvement in results compared to those that were presented in the paper [8], by almost 20%.

We did notice that our model ClinicalICDBERT had a slightly higher precision value than the ClinicalBERT model. We speculate that this may be due to the fact that ICD codes helped in identifying patients that were to be readmitted within the 30 day period. This also means that fewer false positives were detected.

Limitations of using a pre-trained BERT: Most of the drawbacks we faced in improving our results were due to using a pre-trained BERT which was one of the main reasons observed as to why our model never converged. Had we decided to train the BERT model from scratch we would have required more computation due to its size, which comes at a cost. This would require lots more GPUs.

Another problem that we noticed is that because we were using pre-trained embeddings, our model did not converge even after running it for 15 epochs and fell prey to overfitting after the third epoch.

Our ClinicalICDBERT model produced an AUROC of 0.851 which is much better than the AUROC of 0.66 achieved by random forest [8]. We have yet to see the impact of vital signs on our prediction model. However, a useful insight we have gleaned while analyzing the dataset is that stable vital signs are a great measure associated with important clinical consequences. While BERT alone is a powerful model to produce embeddings for text, it has proved to be a powerful model to

produce results for features in both structured and unstructured form. Although our model did not outperform ClinicalBERT significantly, we believe that pre-training the BERT model from scratch on both features instead of using pre-trained embeddings could increase performance by a margin. A major limitation of not pre-training our model from scratch was due to limited computational resources. A model like ClinicalBERT took 2 days to train with a lot of computational resources. Even after using all those resources the model began to overfit after 3 epochs, which confirmed our analysis and experimentations that the pre-trained embeddings of clinicalBERT may not give us much performance boost.

Many of the LM that were pretrained from scratch performed much better than the ones that were simply fine-tuned. even if keeping the architecture the same. Example like bio-bert (see paper)

6. CONCLUSION

Readmissions are a financial burden and a challenge for hospitals. The ability to predict future readmissions can facilitate preventive action. Prior work has explored various structured predictors and machine learning algorithms to predict the risk of readmissions. Clinical notes contain rich data, which is unexploited in predictive modeling compared to structured data. In this work, we are explored a readmission prediction model based on pre-trained clinical notes embeddings and ICD code embeddings. We concatenated both embeddings and fine-tuned our model to predict readmission. Our main contribution was combining structured and unstructured data and using the BERT model for prediction. Our approach performed much better than the machine learning models, achieving an AUROC of 0.851 and a AUPRC of 0.83 but did not improve on results achieved by ClinicalBERT significantly. Readmission predictions can be further improved by training the BERT model from scratch on the input features so as to ensure the model learns better. Readmission prediction is a world-wide challenge being researched by many healthcare institutions. This issue is even more important and challenging if we evaluate hospital and providers based on how patient condition improves over time. These results only highlight the possibility that the concatenation of ICD codes with clinical notes can improve results and create a more robust model.

6.1 Future work

For future work, we can train the BERT model from scratch with the input features instead of using pre-trained embeddings which we saw does not improve the training loss but rather makes it worse over time and also causes overfitting. In addition, MIMIC being a dataset with rich temporal data gives us more opportunity to improve the features we can select to better represent the patient condition. I do believe that incorporating vital signs data for a patient with clinical notes and ICD codes will further improve the results.

As part of future research we can include some descriptive statistics showing the percent of patient records that include vital signs data, ICD codes and clinical notes. As well as those records,

that have all three features for each record. We can also compute, track and use information that shows how many times a patient has been previously been re-admitted (for e.g. in the preceding year) according to the records. This data could also be presented in a time-series format with the dates of prior readmissions.

Additionally, generating results for each feature separately, such as, vital signs data only, ICD codes only, clinical notes only, using the new model ClinicalICDBERT could give us a better idea of the importance and information each feature brings to the model. For our work, we proposed to incorporate the vital signs data but we faced issues with the complexity of the data and a more comprehensive analysis of the features is required to understand how to incorporate these feature in our model. Therefore paying attention to ML features that prior work have found to be important details for vital signs data would be a valid and sensible approach.

While we simply deployed the Binary Cross Entropy(BCE) loss to compute our metrics, modifying the loss function by looking at papers that have used different loss functions, could produce better or optimized results for the metrics that we have used in this work such as AUROC, F1, precision and recall.

Lastly, while our model ClinicalICDBERT, uses both clinical notes and ICD codes we could also experiment by re-crafting previous models such as ClinicalBERT or ML models to produce results for each of the three features separately that is vital signs data only, ICD codes only and clinical notes only. This would give a more comprehensive comparison of the results we get to the results they claim they have gotten on these features, giving us a deeper insight on how we could improve on the implementation details.

REFERENCES

[1] R. Mohammadi, S. Jain, A. T. Namin, M. S. Heller, R. Palacholla, S. Kamarthi, and B. Wallace, "Predicting unplanned readmissions following a hip or knee arthroplasty: Retrospective observational study," JMIR medical informatics, vol. 8, no. 11, p. e19761, 2020.

[2] X. Liu, Y. Chen, J. Bae, H. Li, J. Johnston, and T. Sanger, "Predicting heart failure readmission from clinical notes using deep learning," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2642–2648, IEEE, 2019.

[3] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," Journal of biomedical informatics, vol. 56, pp. 229–238, 2015.

[4] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm, "Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk," Scientific reports, vol. 10, no. 1, pp. 1–10, 2020.

[5] W. Boag, D. Doss, T. Naumann, and P. Szolovits, "What's in a note? unpacking predictive value in clinical note representations," AMIA Summits on Translational Science Proceedings, vol. 2018, p. 26, 2018.

[6] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," arXiv preprint arXiv:1904.05342, 2019.

[7] O. K. Nguyen, A. N. Makam, C. Clark, S. Zhang, B. Xie, F. Velasco, R. Amarasingham, and E. A. Halm, "Vital signs are still vital: instability on discharge and the risk of post-discharge adverse outcomes," Journal of general internal medicine, vol. 32, no. 1, pp. 42–48, 2017.

[8] R. Assaf and R. Jayousi, "30-day hospital readmission prediction using mimic data," in 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–6, IEEE, 2020.

[9] K. Huang, A. Singh, S. Chen, E. T. Moseley, C.-y. Deng, N. George, and C. Lindvall, "Clinical xlnet: modeling sequential clinical notes and predicting prolonged mechanical ventilation," arXiv preprint arXiv:1912.11975, 2019.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, pp. 5998–6008, 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," Advances in neural information processing systems, vol. 32, 2019.

[13] J. Song, M. Hobensack, K. H. Bowles, M. V. McDonald, K. Cato, S. C. Rossetti, S. Chae, E. Kennedy, Y. Barrón, S. Sridharan, et al., "Clinical notes: An untapped opportunity for improving risk prediction for hospitalization and emergency department visit during home health care," Journal of Biomedical Informatics, vol. 128, p. 104039, 2022.

[14] D. J. Feller, J. Zucker, M. T. Yin, P. Gordon, and N. Elhadad, "Using clinical notes and natural language processing for automated hiv risk assessment," Journal of acquired immune deficiency syndromes (1999), vol. 77, no. 2, p. 160, 2018.

[15] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," Journal of the American Medical Informatics Association, vol. 24, no. 1, pp. 198–208, 2017.

[16] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches," Medical care, pp. S106–S113, 2010.

[17] T. Kenzaka, M. Okayama, S. Kuroki, M. Fukui, S. Yahata, H. Hayashi, A. Kitao, D. Sugiyama, E. Kajii, and M. Hashimoto, "Importance of vital signs to the early diagnosis

and severity of sepsis: association between vital signs and sequential organ failure assessment score in patients with sepsis," Internal Medicine, vol. 51, no. 8, pp. 871–876, 2012.

[18] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, and J. Wang, "Automatic icd code assignment of chinese clinical notes based on multilayer attention birnn," Journal of biomedical informatics, vol. 91, p. 103114, 2019.

[19] R. Rutledge, D. B. Hoyt, A. B. Eastman, M. J. Sise, T. Velky, T. Canty, T. Wachtel, and T. M. Osler, "Comparison of the injury severity score and icd-9 diagnosis codes as predictors of outcome in injury: analysis of 44,032 patients," Journal of Trauma and Acute Care Surgery, vol. 42, no. 3, pp. 477–489, 1997.

[20] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage, "Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors," Medical care, vol. 43, no. 5, pp. 480–485, 2005.

[21] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, C. deFilippi, S. R. Steinhubl, and W. F. Stewart, "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2530–2533, IEEE, 2015.

[22] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," IEEE journal of biomedical and health informatics, vol. 22, no. 5, pp. 1589–1604, 2017.

[23] Z. Alyafeai, M. S. AlShaibani, and I. Ahmad, "A survey on transfer learning in natural language processing," arXiv preprint arXiv:2007.04239, 2020.

[24] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials, pp. 15–18, 2019.

[25] A. Rao, E. Barrow, S. Vuik, A. Darzi, and P. Aylin, "Systematic review of hospital readmissions in stroke patients," Stroke research and treatment, vol. 2016, 2016.

[26] S. Howell, M. Coory, J. Martin, and S. Duckett, "Using routine inpatient data to identify patients at risk of hospital readmission," BMC Health Services Research, vol. 9, no. 1, pp. 1–9, 2009.

[27] H. Jasti, E. M. Mortensen, D. S. Obrosky, W. N. Kapoor, and M. J. Fine, "Causes and risk factors for rehospitalization of patients hospitalized with community-acquired pneumonia," Clinical infectious diseases, vol. 46, no. 4, pp. 550–556, 2008.

[28] A. Gruneir, I. A. Dhalla, C. van Walraven, H. D. Fischer, X. Camacho, P. A. Rochon, and G. M. Anderson, "Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm," Open Medicine, vol. 5, no. 2, p. e104, 2011.

[29] K. K. Lee, J. Yang, A. F. Hernandez, A. E. Steimle, and A. S. Go, "Post-discharge follow-up characteristics associated with 30-day readmission after heart failure hospitalization," Medical care, vol. 54, no. 4, p. 365, 2016.

[30] P. K. Lindenauer, S.-L. T. Normand, E. E. Drye, Z. Lin, K. Goodrich, M. M. Desai, D. W. Bratzler, W. J. O'Donnell, M. L. Metersky, and H. M. Krumholz, "Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia," Journal of Hospital Medicine, vol. 6, no. 3, pp. 142–150, 2011.

[31] E. H. Bradley, L. Curry, L. I. Horwitz, H. Sipsma, J. W. Thompson, M. Elma, M. N. Walsh, and H. M. Krumholz, "Contemporary evidence about hospital strategies for reducing 30-day readmissions: a national study," Journal of the American College of Cardiology, vol. 60, no. 7, pp. 607–614, 2012.

[32] A. Artetxe, A. Beristain, and M. Grana, "Predictive models for hospital readmission risk: A systematic review of methods," Computer methods and programs in biomedicine, vol. 164, pp. 49–64, 2018.

[33] K. Teo, C. W. Yong, J. H. Chuah, B. P. Murphy, and K. W. Lai, "Early detection of readmission risk for decision support based on clinical notes," Journal of Medical Imaging and Health Informatics, vol. 11, no. 2, pp. 529–534, 2021.

[34] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. Waljee, "Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review," bmj, vol. 369, 2020.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, pp. 3111–3119, 2013.

[36] M. Kholghi, L. De Vine, L. Sitbon, G. Zuccon, and A. Nguyen, "The benefits of word embeddings features for active learning in clinical information extraction," arXiv preprint arXiv:1607.02810, 2016.

[37] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[38] J. Libovickỳ, R. Rosa, and A. Fraser, "How language-neutral is multilingual bert?," arXiv preprint arXiv:1911.03310, 2019.

[39] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare mimic datasets," arXiv preprint arXiv:1710.08531, 2017.

[40] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in International conference on machine learning, pp. 1188–1196, PMLR, 2014.

[41] S. N. Golmaei and X. Luo, "Deepnote-gnn: predicting hospital readmission using clinical notes and patient network," in Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1–9, 2021.