Heliyon 10 (2024) e25960

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A two-stream deep model for automated ICD-9 code prediction in an intensive care unit

Mustafa Arda Ayden [a], Mehmet Eren Yuksel [b,*], Seniha Esen Yuksel Erdem [a]

[a] Department of Electrical and Electronics Engineering, Hacettepe University, Ankara, 06800, Türkiye
[b] Surgical Intensive Care Unit, Ankara Etlik City Hospital, Ankara, 06170, Türkiye

ARTICLE INFO

Keywords:
Deep learning
Electronic health records
MIMIC-III
ICD-9
ICU

ABSTRACT

Assigning medical codes for patients is essential for healthcare organizations, not only for billing purposes but also for maintaining accurate records of patients' medical histories and analyzing the outputs of certain procedures. Due to the abundance of disease codes, it can be laborious and time-consuming for medical specialists to manually assign these codes to each procedure. To address this problem, we discuss the automatic prediction of ICD-9 codes, the most popular and widely accepted system of medical coding. We introduce a two-stream deep learning framework specifically designed to analyze multi-modal data. This framework is applied to the extensive and publicly available MIMIC-III dataset, enabling us to leverage both numerical and text-based data for improved ICD-9 code prediction.

Our system uses text representation models to understand the text-based medical records; the Gated Recurrent Unit (GRU) to model the numerical health records; and fuses these two streams to automatically predict the ICD-9 codes used in the intensive care unit. We discuss the preprocessing and classification methods and demonstrate that our proposed two-stream model outperforms other state-of-the-art studies in the literature.

1. Introduction

To accurately document surgical procedures and diagnoses, the World Health Organization (WHO) has developed the International Classification of Diseases, Ninth Revision (ICD-9) coding system. This system standardizes the surgical codes and descriptions, helps the billing processes, and supports public health surveillance and research efforts. ICD-9 codes consist of three to five digits. The first three digits represent the main category of the diagnosis or procedure, while the fourth and fifth digits provide additional detailed information. These codes are essential for tracking the results of procedures over time, and they can help healthcare providers identify opportunities for improvement. For instance, physicians might determine ways to lower the risk of problems by examining the frequency of certain surgical operations and their outcomes. However, most medical coding is presently done manually, which is prone to errors and labor-intensive [1].

To help improve the healthcare industry, medical code prediction based on test results has become an important area of research in machine learning. By automating the process, doctors can make more accurate and timely diagnoses, and increase the quality of surgical care. In this work, we analyze the latest best practices for medical code prediction and propose an algorithm to predict the medical codes of both the procedures applied to a patient and the diagnoses made by the doctors. We use the publicly available

* Corresponding author.
E-mail address: doctormehmeteren@yahoo.com (M.E. Yuksel).

https://doi.org/10.1016/j.heliyon.2024.e25960
Received 9 January 2024; Received in revised form 30 January 2024; Accepted 5 February 2024

Available online 8 February 2024

MIMIC-III healthcare dataset [2] collected in the patient recovery process from 58,576 adults in an intensive care unit. In this dataset, there are both numerical test results as well as unstructured written documents. While most of the work in the literature for ICD-9 code prediction is performed on unstructured text data, in this work, we are proposing a two-stream method that investigates both the unstructured texts and the time series of numerical features based on transformer methods, comparing them to the state-of-the-art works in the literature.

In this study, our goal is to increase the accuracy and timeliness of diagnoses and procedures by automating the prediction of ICD-9 codes. The scope of the work involves the separate examination of text-based data and time-dependent numerical data within the MIMIC-III dataset. Later, deep learning-based methods that utilize both data types are employed. Different deep learning methods are trained and tested with the dataset for both data types. By employing deep learning techniques, the study achieves the best results for both data types. These successful methods are then combined to form an ensemble model, which is adopted as the final approach. Consequently, the developed method is capable of effectively estimating the ICD-9 codes of patients by using both text-based and numerical data as input.

We achieve better results in medical code prediction by utilizing deep learning techniques for both types of data and combining them in an ensemble model. Our findings indicate that when text-based methods are integrated with numerical data approaches, there is a significant improvement in accuracy, which highlights the effectiveness of our two-stream model.

The findings revealed that the results obtained from text-based data were more successful than those from time series data. However, when the text-based estimation methods were combined with methods using numerical data, the overall model's success was found to increase significantly. To validate this outcome, the same approach was applied to the most successful text-based methods reported in the literature, which also demonstrates that the performance of these methods can be enhanced through a similar ensemble. Overall, the two-stream model achieved better results by leveraging the strengths of both text-based and numerical data, leading to improved ICD-9 code estimation.

2. Dataset and preprocessing

The MIMIC-III dataset [2] used in this study contains health data of patients who were treated in the intensive care unit (ICU) of the Beth Israel Deaconess Medical Center in the United States between 2001 and 2012. During MIMIC's data collection period, two different critical care information systems were in place, making this dataset diverse in this sense. The dataset comprises clinical notes, time-dependent numeric data, and time-independent numeric data from 58,576 patient admissions, of which 53,423 are adults. The median age of the patients in the dataset is 65.68 years, and the median length of hospital stay is 7.08 days. The hospital mortality rate is 10.49%. Among adult patients, the first admission rate is 83%, and there is data on 38,425 different adult patients. For this study, only adult patient data were used, and if a patient had more than one admission, their first admission was selected. We focus on ICU data because ICU measurements are more common in the dataset we used. Additionally, the methods we compare are also designed for intensive care data, further ensuring a fair comparison. However, it is worth noting that the algorithms we propose can be generalized to other ICD code prediction areas beyond the ICU with further experiments.

To predict ICD-9 codes using both text and time-dependent numerical data, we utilize the *noteevents*, *chartevents*, *inputevents* and *labevents* tables as input. The *noteevents* table contains all the text-based data obtained during the hospitalization process, while the other tables contain numerical data collected over a specific period. In line with previous studies in the literature, only the patient discharge notes in the *noteevents* table were used as text-based data. These discharge notes are also the only documents used by clinical coders to assign medical codes. A short example of these discharge summaries can be found in Fig. 1. The discharge summaries are typically much longer.

For outputs in training our classifiers, we utilize the *diagnoses_icd* and *procedures_icd* tables as ICD-9 code labels of patients. ICD-9 codes consist of 3 to 5 digits. The format of the 5-digit code is XXX.XX, where the first three digits are used for the disease category, and the last two digits are used for the etiology of the disease in two sub-breaks. In other words, codes with 3 digits represent the main diseases, while the 4th and 5th digits constitute the sub-branches of the main diseases. For example, the ICD-9 code 290 is used to represent dementia, the code 290.1 to represent early dementia, and the code 290.12 to represent early delusional dementia.

Inevitably, the electronic health records that comprise the MIMIC-III dataset have some limitations. These include high missing rates since not all tests are administered to all patients, multidimensionality of each patient data, high noise, and random errors. Most

167853 Admission Date: [**2151-7-16**] Discharge Date: [**2151-8-4**]

Service: ADDENDUM:

RADIOLOGIC STUDIES: Radiologic studies also included a chest CT, which confirmed cavitary lesions in the left lung apex consistent with infectious process/tuberculosis. This also moderate-sized left pleural effusion.

HEAD CT: Head CT showed no intracranial hemorrhage or mass effect, but old infarction consistent with past medical history.

ABDOMINAL CT: Abdominal CT showed lesions of T10 and sacrum most likely secondary to osteoporosis. These can be followed by repeat imaging as an outpatient.

[**First Name8 (NamePattern2) **] [**First Name4 (NamePattern1) 1775**] [**Last Name (NamePattern1) **], M.D. [**MD Number(1) 1776**]

Dictated By:[**Hospital 1807**] MEDQUIST36

D: [**2151-8-5**] 12:11 T: [**2151-8-5**] 12:21

Fig. 1. Sample Discharge Summary from MIMIC-III Dataset.

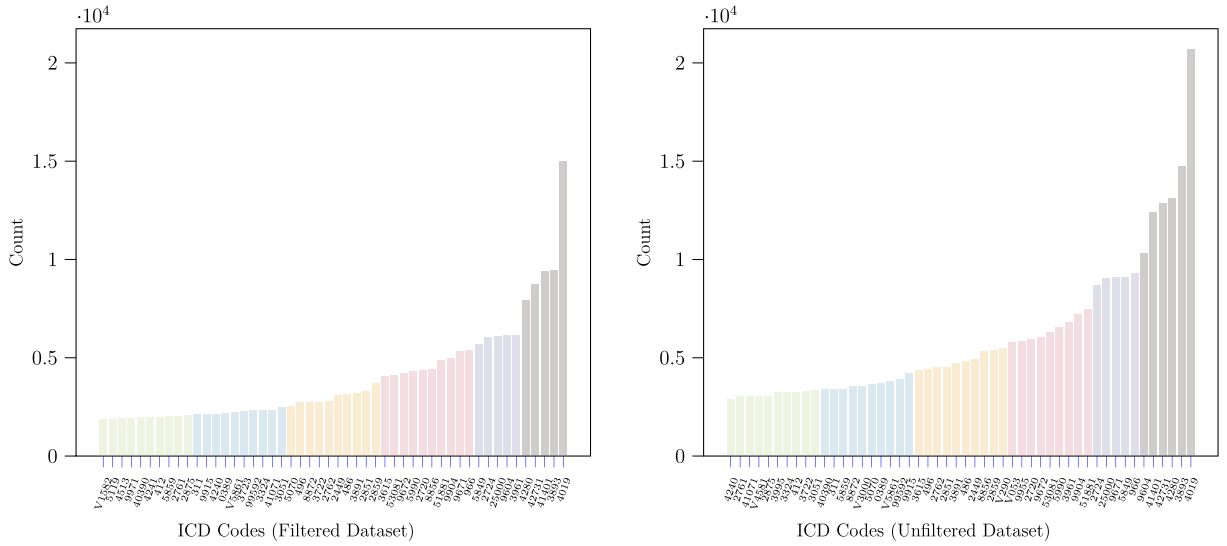

Fig. 2. The 50 Most Common ICD Code Counts in both Filtered and Unfiltered Datasets.

importantly, each patient has a large number of parameters, and different parameters are needed for diagnosing different diseases. This leads to many missing data for each patient, which must be taken into account when analyzing electronic health data. Below, we describe how to process the dataset to handle these problems and make it ready for classification. After the preprocessing, we end up with three datasets, namely Dataset A, B, and C. Dataset A uses most of the available data and is our preferred dataset. The last two datasets were formed to be on the same scale as the other works in the literature for a fair comparison.

Dataset A: In MIMIC-III, some patients are in the text dataset but not in the numeric dataset. With this in mind, patients in the dataset were filtered and patients who exist in both the numeric and text datasets were taken. For label data, the 50 most common ICD codes in the dataset were used to create a fair comparison environment with the studies in the literature. Patients who did not have any of these 50 most common ICD codes were also excluded from the dataset. The resulting list of codes is given in Table 2, and statistics of the 50 most common ICD codes in both filtered and unfiltered (raw) datasets are also available in Fig. 2.

The resulting filtered, final dataset contains data from 33330 different patients. This dataset is referred to as Dataset A for the remainder of the study. For each experiment performed, the dataset was randomly divided into a training set, validation set, and test set, with ratios of 0.7, 0.15, and 0.15, respectively, and the trainings were carried out. The same training, validation, and test sets were used for each experiment to create a fair comparison environment.

Dataset B: Xie et al. [3] arranged the discharge reports into training, validation, and test datasets based on 8066, 1728, and 1729 patient referrals from the MIMIC-III dataset, respectively. Afterward, this dataset was used in various state-of-the-art studies including [4–7]. In this study, this dataset is referred to as Dataset B.

Dataset C: To compare the proposed method with other studies, it was necessary to conduct some experiments with Dataset B for the proposed methods. However, since there is no numerical record of some patient applications in the Dataset, they were excluded from our study. For the training, validation, and test datasets, missing patient referrals were 2716, 395, and 400, respectively. Excluding these patient referrals, the training set consisted of 5350 patient referrals, the validation set consisted of 1333, and the test dataset consisted of 1329 patient referrals. This filtered dataset is named Dataset C in the rest of the study. In other words, Dataset C can be seen as a sub-dataset of Dataset B.

For numerical feature selection, the most comprehensive feature set in [8] was used. In this sub-dataset of the MIMIC-III, there are 136 time-dependent features from more than 20,000 patient referrals. These attributes are included in Table 1. Some attribute names in the dataset can appear in both uppercase and lowercase letters, so the names in the table are left as they are in the dataset. Also, since there are attributes with the same name, some attributes are included with a single name. The features in the table were handled for 24 hours from the patient's admission and were sampled in one-hour slices. Therefore, it is possible to say that the obtained numerical data are time-dependent and 136x24 in size. Numerical tables in the MIMIC-III dataset have some problems due to their structure, as described. The main problems are high rates of missing data and data representing the occurrence of the same attribute in more than one table. These problems need to be taken into account in the pre-processing stage. Hence, attributes from more than one table are combined into a single attribute, as in the study of Purushotham et al. [8]. Values in the dataset are normalized for each feature. In addition, for missing data, if that attribute is included in a patient's application at any time, forward and backward filling processes are applied, respectively. If a patient referral has no record of that attribute for any period, these missing fields are filled with zeros.

For text data, only discharge reports from the text-based reports in the MIMIC-III dataset were used as in other studies. For data preprocessing, the date, time, and special characters in each discharge report were deleted and all letters were reduced.

Table 1
Numerical Features Used from the MIMIC-III Dataset with their Original Names (i.e. with both uppercase and lowercase letters as given in the dataset).

Albumin 5	Pre-Admission	Bisacodyl
Fresh Frozen Plasma	TF Residual	Docusate Sodium
Lorazepam (Ativan)	urinary output sum	Humulin-R Insulin
Calcium Gluconate	HEMATOCRIT	Metoprolol Tartrate
Midazolam (Versed)	PLATELET	Pantoprazole
Phenylephrine	HEMOGLOBIN	ArterialBloodPressurediastolic
Furosemide (Lasix)	MCHC	ArterialBloodPressuremean
Hydralazine	MCH	RespiratoryRate
Norepinephrine	MCV	AlarmsOn
Magnesium Sulfate	RED BLOOD CELLS	MinuteVolumeAlarm-Low
Nitroglycerin	RDW	Peakinsp.Pressure
Insulin - Regular	CHLORIDE	PEEPset
Morphine Sulfate	ANION GAP	MinuteVolume
Potassium Chloride	CREATININE	GLUCOSE
Gastric Meds	MAGNESIUM, TOTAL	weight
D5 1/2NS	CALCIUM	height
LR	PHOSPHATE	glucose
Solution	INR(PT)	spo2 peripheral
Sterile Water	PT	arterial pressure mean
Piggyback	PTT	diastolic blood pressure mean
OR Crystalloid Intake	LYMPHOCYTES	ie ratio mean
PO Intake	MONOCYTES	fio2
GT Flush	NEUTROPHILS	body temperature
KCL (Bolus)	BASOPHILS	heart rate
Magnesium Sulfate (Bolus)	EOSINOPHILS	systolic blood pressure abp mean
epinephrine	PH	gcseyes
vasopressin	BASE EXCESS	gcsmotor
dopamine	CALCULATED TOTAL CO2	gcsverbal
midazolam	PCO2	SkinCare
fentanyl	SPECIFIC GRAVITY	RespAlarm-High
propofol	LACTATE	SpO2DesatLimit
Gastric Tube	ALANINE AMINOTRANSFERASE (ALT)	PulmonaryArteryPressurediastolic
Stool Out Stool	ASPARATE AMINOTRANSFERASE (AST)	TidalVolume(set)
Urine Out Incontinent	ALKALINE PHOSPHATASE	PulmonaryArteryPressuresystolic
Ultrafiltrate	ALBUMIN	HeartRateAlarm-Low
Fecal Bag	pao2	Glucosefingerstick
Chest Tube #1	serum urea nitrogen level	O2Flow
Chest Tube #2	white blood cells count mean	PulmonaryArteryPressuremean
Jackson Pratt #1	serum bicarbonate level mean	RespiratoryRate(Set)
OR EBL	sodium level mean	CentralVenousPressure
potassium level mean	bilirubin level	MeanAirwayPressure
hgb	chloride	TidalVolume(observed)
peep	Aspirin	MinuteVolumeAlarm-High
Packed Red Blood Cells		

3. Literature review

In the quest for automating the ICD code prediction, traditional methods employed in this field rely on scoring systems, specifically SAPS [9], SOFA [10] and APACHE [11], which are expressed through mathematical formulas. These methods involve calculating scores based on a predefined set of attributes determined by physicians. These scores help healthcare professionals predict the survival status of patients in the intensive care unit by assessing the severity of their health conditions.

With the increasing success of machine learning algorithms, several methods have been proposed, offering superior results compared to traditional approaches. In a study conducted by Purushotham et al. [8] various machine learning and deep learning methods used in this field were compared, demonstrating their superior performance over traditional scoring methods. The study employed a GRU-based model as a deep learning method to predict outcomes from time-dependent data. To conduct the analysis, three distinct sub-datasets were created using the MIMIC-III dataset, each containing different numerical features. The tasks examined for each dataset included estimating hospital mortality, short-term mortality, long-term mortality, hospital stay, and ICD code prediction. The first dataset consisted of 17 features used in the SAPS scoring system, the second dataset contained 20 raw features from which the SAPS scoring system features were derived, and the third dataset incorporated 136 raw features irrespective of scoring systems. The results revealed that deep learning-based methods outperformed other machine learning and scoring methods across all tasks, particularly in experiments involving a large number of features and raw data. Consequently, we also employ the most prevalent 136 raw features from the numerical data for the estimation task.

Datasets that include electronic health records encompass more than just numerical information. Within these datasets, one can often encounter clinical notes obtained during hospital admission and stay. Notably, a significant portion of the MIMIC-III dataset primarily comprises these notes. As a result, the advancements in deep learning techniques within the field of natural language processing have led to numerous studies in this domain [12,13]. Consequently, estimation approaches based on clinical health

Table 2
50 Most Common ICD-9 Code Descriptions.

ICD9 Code	Count	Short Name	Long Name	Type
4019	15030	Hypertension NOS	Unspecified essential hypertension	Diagnosis
3893	9476	Venous cath NEC	Venous catheterization, not elsewhere classified	Procedure
41401	9393	Crnry athrscl natve vssl	Coronary atherosclerosis of native coronary ar...	Diagnosis
42731	8758	Atrial fibrillation	Atrial fibrillation	Diagnosis
4280	7933	CHF NOS	Congestive heart failure, unspecified	Diagnosis
3961	6177	Extracorporeal circulat	Extracorporeal circulation auxiliary to open heart surgery	Procedure
9604	6165	Insert endotracheal tube	Insertion of endotracheal tube	Procedure
25000	6113	DMII wo cmp nt st uncntr	Diabetes mellitus without mention of complicat...	Diagnosis
2724	6034	Hyperlipidemia NEC/NOS	Other and unspecified hyperlipidemia	Diagnosis
5849	5695	Acute kidney failure NOS	Acute kidney failure, unspecified	Diagnosis
966	5388	Entral infus nutrit sub	Enteral infusion of concentrated nutritional substances	Procedure
9671	5340	Cont inv mec ven <96 hrs	Continuous invasive mechanical ventilation for less than 96 consecutive hours	Procedure
9904	4967	Packed cell transfusion	Transfusion of packed cells	Procedure
51881	4904	Acute respiratry failure	Acute respiratory failure	Diagnosis
8856	4447	Coronar arteriogr-2 cath	Coronary arteriography using two catheters	Procedure
2720	4388	Pure hypercholesterolem	Pure hypercholesterolemia	Diagnosis
5990	4312	Urin tract infection NOS	Urinary tract infection, site not specified	Diagnosis
9672	4249	Cont inv mec ven 96+ hrs	Continuous invasive mechanical ventilation for 96 consecutive hours or more	Procedure
53081	4138	Esophageal reflux	Esophageal reflux	Diagnosis
3615	4086	1 int mam-cor art bypass	Single internal mammary-coronary artery bypass	Procedure
2859	3738	Anemia NOS	Anemia, unspecified	Diagnosis
2851	3310	Ac posthemorrhag anemia	Acute posthemorrhagic anemia	Diagnosis
3891	3191	Arterial catheterization	Arterial catheterization	Procedure
486	3134	Pneumonia, organism NOS	Pneumonia, organism unspecified	Diagnosis
2449	3086	Hypothyroidism NOS	Unspecified acquired hypothyroidism	Diagnosis
2762	2818	Acidosis	Acidosis	Diagnosis
3722	2768	Rflx sym dystrph lwr lmb	Reflex sympathetic dystrophy of the lower limb	Diagnosis
8872	2749	Dx ultrasound-heart	Diagnostic ultrasound of heart	Procedure
496	2733	Chr airway obstruct NEC	Chronic airway obstruction, not elsewhere clas...	Diagnosis
5070	2567	Food/vomit pneumonitis	Pneumonitis due to inhalation of food or vomitus	Diagnosis
3051	2485	Tobacco use disorder	Tobacco use disorder	Diagnosis
41071	2358	Subendo infarct, initial	Subendocardial infarction, initial episode of ...	Diagnosis
3324	2348	Closed bronchial biopsy	Closed [endoscopic] biopsy of bronchus	Procedure
99592	2330	Severe sepsis	Severe sepsis	Diagnosis
3723	2266	Conjunctivitis NOS	Conjunctivitis, unspecified	Diagnosis
V5861	2231	Long-term use anticoagul	Long-term (current) use of anticoagulants	Diagnosis
0389	2203	Septicemia NOS	Unspecified septicemia	Diagnosis
4240	2148	Mitral valve disorder	Mitral valve disorders	Diagnosis
9915	2120	Parent infus nutrit sub	Parenteral infusion of concentrated nutritional substances	Procedure
311	2116	Depressive disorder NEC	Depressive disorder, not elsewhere classified	Diagnosis
2875	2091	Thrombocytopenia NOS	Thrombocytopenia, unspecified	Diagnosis
2761	2039	Hyposmolality	Hyposmolality and/or hyponatremia	Diagnosis
5859	2037	Chronic kidney dis NOS	Chronic kidney disease, unspecified	Diagnosis
412	2004	Old myocardial infarct	Old myocardial infarction	Diagnosis
4241	2002	Partial esophagectomy	Partial esophagectomy	Procedure
40390	1976	Hy kid NOS w cr kid I-IV	Hypertensive chronic kidney disease, unspecifi...	Diagnosis
9971	1929	Therapeu plasmapheresis	Therapeu plasmapheresis	Procedure
4513	1920	Sm bowel endoscopy NEC	Other endoscopy of small intestine	Procedure
5119	1885	Biliary tr dx proc NEC	Other diagnostic procedures on biliary tract	Procedure
V1582	1873	History of tobacco use	Personal history of tobacco use	Diagnosis

reports have been developed. The concept of the Clinical Bert model has been mentioned in studies such as the studies of Alsentzer et al. [14] and Huang et al. [15]. Clinical BERT is a customized BERT [16] model specifically tailored for clinical applications. BERT itself is a deep learning model based on transformer encoders, which vectorizes words or text phrases by considering their contextual properties. Unlike other deep learning-based natural language processing techniques, this model processes word sequences bidirectionally, both left-to-right and right-to-left, aiming to improve the prediction of word relationships. Since the original BERT model was trained on a Wikipedia dataset that did not adequately represent medical sentences, Alsentzer et al. conducted model training using clinical texts instead of relying on the pre-trained BERT model. Huang et al. used the Clinical BERT model to estimate 30-day hospital readmission. Their study utilized various clinical notes, including evacuation notes from the MIMIC-III dataset and ECG and radiology reports gathered during intensive care unit stays. This approach enables the estimation of hospitalization time upon patient discharge using text-based data acquired at any point during the intensive care hospitalization period. When evaluating the study's results, it becomes evident that the Clinical BERT model outperforms BERT, Bag-of-words, and BI-LSTM methods in terms of producing more accurate outputs.

In a study published by Jin et al. [17], both textual and numerical data were used simultaneously. In this study, in which a multimodal deep learning method is proposed, the task of hospital mortality is discussed. An LSTM model is used for numerical data, and a deep learning method based on Doc2VecC [18] is used for text-based data. In the training phase, these two models are

connected and the vectors obtained as the output of the two models are combined into a single vector and given as input to the fully connected layer. Thus, both numerical and textual data were taught at the same time. Time-dependent data sampled in two-hour periods for the first 48 hours for each patient's 17 features were given as input to the LSTM-based deep learning network used for numerical data. In the study of Jin et al., a medical NER-based service was used to be able to itemize sentences in the deep learning network used for learning text-based data [19]. Lexical itemized texts are given as input to the Doc2VecC model. The Doc2VecC model is a method for extracting a vector representation of all text. In the training of this model, each word output is given as input as well as randomly selected words as well as neighboring words. As a side effect of using random vectors, it is aimed that the averages of the word vectors represent the document vector. Considering that the Doc2Vec model is trained only with neighboring words, it is possible to say that the Doc2VecC model is more successful in representing the document vector. In addition, the model requires less effort for new data that has not yet been trained. This method, which is generally used for sentiment analysis, document classification, and semantic relatedness tasks, was used for the representation of clinical texts in the study. As a result of this study using the MIMIC-III dataset, it has been shown that such a multimodal deep learning method gives a 2% more successful result in estimating hospital mortality compared to other uniform (numerical or textual) methods.

Vu et al. proposed a deep learning model with a tag attention mechanism for estimating ICD code from text data [5]. In this study, Word2Vec was used as the basic natural language processing method. In addition, a BiLSTM model that takes Word2Vec word vectors as input is positioned to obtain the contextual properties of the words in the clinical notes. For the attention mechanism, the tag attention layer is applied after the BiLSTM layer, and a structure that can generate a different vector for each ICD code tag to be used as output is obtained. The label attention layer is based on the structured self-attention mechanism proposed by Lin et al. [20]. Based on the 50 most common ICD-9 codes, this study seems to yield more successful results than other attentional networks.

Reys et al. have created various classifiers based on the Word2Vec model for the task of estimating ICD code from text data in [21]. They created a dictionary with the text-based reports they obtained from the MIMIC-III dataset and completed the training of the Word2Vec model. The report vectors obtained with Word2Vec were tested with three different classifiers. These classifiers are a convolutional neural network-based classifier, a convolutional neural network-based classifier with an attention mechanism, and a GRU-based classifier.

Yang et al. proposed a deep learning method based on the Longformer method, powered by the *Knowledge Enhanced PrompT (KEPT)* method [7]. They explained that they chose a Longformer-based method as the natural language processing (NLP) method because the texts in the MIMIC-III dataset were too large to be fully represented by other methods. In this method, while giving the texts as input to the Longformer model, they also gave the segmented versions of the ICD code classes as input. Therefore, they aimed for the model to learn the ICD codes along with the texts in the MIMIC-III dataset.

As a result of the literature review, although there are studies that make various types of healthcare prediction using numerical methods, these studies generally focused on other estimation tasks rather than ICD code estimation. The study of Purothom et al. is one of the most comprehensive examples of them.

Text-based methods are generally diversifying in ICD code estimation studies. The latest increase in the number of text-based methods in ICD code estimation studies can be evaluated as a result of the massive progress in NLP literature. In different studies, deep learning models using various NLP methods have been proposed and their superiority over each other has been revealed. As mentioned above, the studies of Huang et al., Vu et al., Reys et al., and Yang et al. are examples of these studies using only text-based data. Table 3 shows the comparison of these and the other literature studies evaluated in this work in terms of target, data type used, and NLP method.

Table 3
Comparison of studies in the literature.

Work	Year	Estimation Task					Data Type			NLP Method Used				
		ICD Code	ICD Code Group	Mortality	Length of Stay	Hospital Readmission	Structured Numerical	Unstructured Text	Both	Doc2Vec	Doc2VecC	Bert	Word2Vec	Longformer / KEPT
S. Purushotham [8]	2017		✓	✓	✓		✓							
Mengqi Jin [17]	2018			✓					✓		✓			
J. Mullenbach [4]	2020	✓						✓					✓	
Kexin Huang [15]	2020					✓		✓				✓		
L. Franz [22]	2020	✓	✓				✓	✓				✓		
Thanh Vu [5]	2020	✓						✓					✓	
Arthur D. Reys [21]	2020	✓						✓					✓	
Yang Liu [23]	2021	✓						✓				✓	✓	
Zhichao Yang [7]	2022	✓						✓						✓
Zheng Yuan [6]	2022	✓						✓					✓	
This Work		✓							✓			✓		✓

Our study can be seen as a study that uses both data types for the ICD estimation task and shows that successful results can be obtained when both data types are used. Although some previous works have utilized both data types, similar to the study conducted by Jin et al., they were primarily applied to tasks other than the ICD estimation task. Our study takes a multi-instrumental approach, aligning with Jin et al.'s work. However, there is a key difference: while Jin et al.'s study focused on estimating hospital mortality, our research centered on estimating ICD codes. As a result, Jin et al.'s work addresses a two-class single-label task, whereas our study aims to tackle a multi-class multi-label problem.

4. Method

In this study, both time-dependent numeric data and text-based clinical notes from the MIMIC-III dataset were used to estimate ICD codes in a two-stream network. Initially, these two types of data were handled as separate tasks. Deep learning models that yielded the best results for each task were evaluated as separate streams, one as a text stream, and the other as a numerical stream; intended to be combined in the main task to achieve more robust outcomes. The outputs of these two streams were combined using a fully connected layer. Once the deep learning model for each stream was determined and preprocessing was applied, the entire proposed model was trained. Given that the task involves a multi-label multi-class structure, the binary cross-entropy with logit loss function was used as the error function during the training of each model.

4.1. Text stream

The text stream was trained to handle unstructured texts as in the medical records. For text-based classification, state-of-the-art transformer-based NLP methods were utilized and evaluated. These methods include Clinical BERT [14], Clinical Longformer [24] and KEPT [7] based methods, respectively.

4.1.1. Clinical BERT

Unlike other word representation methods such as Word2Vec [25], the BERT and the Clinical BERT models are context-dependent, generating different vectors for the same words in different contexts. Moreover, they are designed not only for word-level vectors but also for representing sentences and paragraphs. In this study, we aim to represent clinical documents and classify them into ICD-9 codes based on these representations, as the dataset mainly consists of paragraphs with multiple sentences. Consequently, paragraph encodings from the output of the Clinical BERT model are utilized to represent clinical notes. However, the original BERT and Clinical BERT models were designed to handle a maximum input length of 512 tokens for each instance. In contrast, the clinical notes in the MIMIC-III dataset have longer sequences, with mean and median word sequence lengths of 1528 and 1415, respectively.

To address this issue, we split the texts into two sub-chunks, tokenize the first 512 words of each chunk, and use both chunks as input to the Clinical BERT model. The resulting encodings from both chunks are then concatenated, resulting in each text being represented by 1x1536 vectors, as the model has a hidden size of 768 for a single chunk. For the ICD-9 classification, a single fully connected layer is used. Given that each text is represented by a 1536-sized vector and there are 50 classes, the dimensions of the classification layer are 1536x50. The model is trained with Dataset A, which is the most comprehensive dataset in our study.

4.1.2. Clinical longformer

While the BERT model can deliver highly successful outputs for NLP tasks, it faces limitations in processing long strings, such as paragraphs. The standard BERT model and some other versions like Clinical BERT are designed to handle texts with a maximum of 512 tokens. To address longer texts, one approach is to split them into separate chunks and input them into the model. However, using an excessive number of chunks can lead to challenges in effectively representing relatively shorter texts in the dataset.

Longformer [26] is a transformer-based method developed to overcome these issues and process long texts as input effectively. Unlike the BERT model, the Longformer can handle a larger token size, up to 4096 tokens. To enable the use of longer sequences as input, the Longformer model incorporates various attention methods in addition to the standard transformer-based methods. These include sliding window attention, dilated sliding window attention, and global attention.

In this study, for the Longformer-based text classification, the Clinical Longformer model [24] developed for clinical studies was examined and used. The structure of this model is in the same architecture as the basic longformer model and differs in terms of the dataset used. The Clinical Longformer model has been trained with the MIMIC-III dataset which is used in this study. This model is also trained with the Dataset A.

4.1.3. KEPT

To generalize the results obtained in this study, experiments were also carried out using the KEPT method, which has the most successful result in the literature for the top 50 ICD-9 prediction tasks. The KEPT method is another Longformer-based language model proposed by Yang et al. [7] for ICD encoding purposes. In this method, an alternative approach is developed for multi-class tasks. In the developed model, the input comprises descriptions of the ICD codes targeted for estimation, along with the discharge reports. At this point, firstly, the tokens for discharge notes and ICD code descriptions were obtained with the pre-trained Clinical Longformer method. In addition, while ICD code descriptions are given as input, [MASK] token is added to the end of these descriptions, aiming to predict this tokenized item by the language model. In the output, a one-to-one coding-like structure was created by representing the owned ICD codes with "yes" tokens and not owned ICD codes with "no" tokens for each patient application. Thus, in the output, the [MASK] tokens that the language model is estimated as "yes" are obtained as the predicted ICD code, and the tokens that are

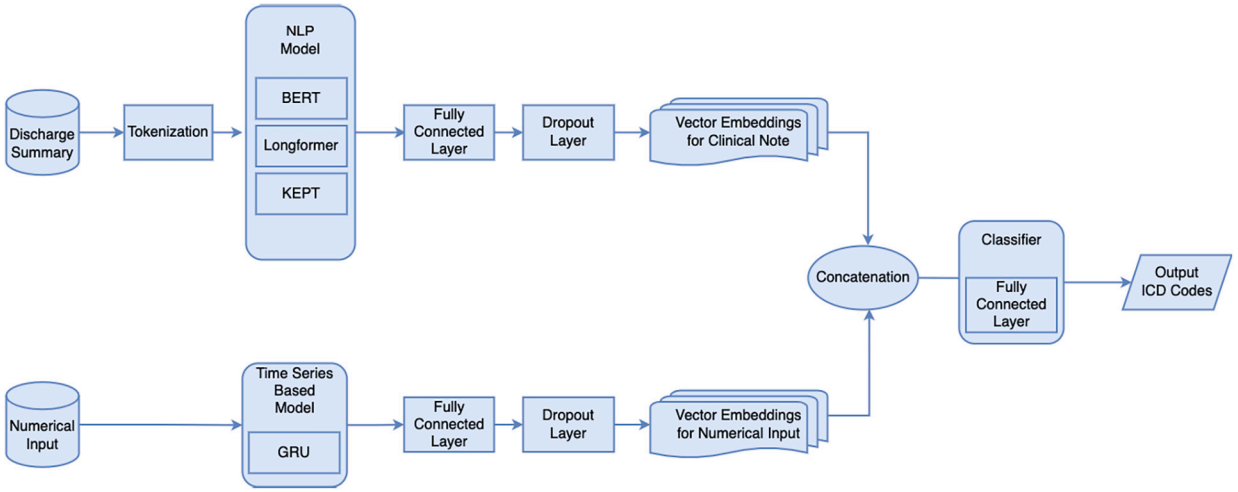

Fig. 3. Ensemble Model Flow Diagram.

predicted as 'no' are obtained as the unpredicted ICD codes. The base KEPT model was trained with Dataset B by Yang et al. However, in this study, while training the KEPT model, the C dataset was used to compare the difference between the KEPT model and the ensemble model fairly. Dataset C is nothing but a subset of Dataset B, including its text data, as mentioned in Section 2.

The above three methods were trained and fine-tuned by using the Adam optimizer with the same learning rate of 0.001 for 15 epochs. BCE with logit loss was used and the states that gave the best validation F1 score for each method were evaluated as the best models. Additionally, these best models for each approach were used independently to create ensemble models.

4.2. Numerical stream

This method was developed for processing the time series of numerical features as input. Since there are 136 features with 24-hour windows, the input dimensions are 24 for each patient. For representing the data which is in time series format, a GRU layer was used in the first layer with 20 features in hidden size and 2 recurrent layers. Using GRU makes it possible to cover time dependency given instances. The output dimension of the layer is 136x20 since there are 136 features in the input and the hidden size is 20. In the second layer, the GRU output is flattened to 2720 to feed the classification layer. For the classification part of the network, two fully connected layers with 100, and 50 (number of ICD-9 codes) units were used respectively. The ReLu activation function was used for the fully connected layers. The model is trained over 750 epochs with a batch size of 64. The Adam optimizer was used with a 1e-3 initial learning rate, which was reduced by half at every 100 epochs. As in models for the text-based classification task, BCE with logit loss was used and the state that gives the best validation F1 score was evaluated as the best model for the ensemble models. The GRU model for representing the time series of numerical features is trained with Dataset A.

4.3. Integrating text and numerical streams

After training the models for text-based and numerical data separately, each model pair was ensembled without their classification layers. Then, a single classification layer was added after jointly concatenating the outputs of each model. Two dropout layers with a probability of $p = 0.4$ were used in the ensemble models: the first one for the concatenation of the text-based model output and the second one for the concatenation of both models. The resulting model can be found in Fig. 3. As an initial network, pre-trained models described in the above sections were used and fine-tuned collectively. For the training process of the ensemble models, the Adam optimizer is used with a learning rate of 1e-5. The models were trained over 30 epochs with a batch size of 8, and the results of the state that achieved the best validation F1 scores were reported as the final ensemble model. The ensemble model is trained with Dataset A for the Clinical BERT and Clinical Longformer-based models and Dataset C for the KEPT-based model. Thus, the results of this study were validated both with Clinical BERT and Clinical Longformer in comprehensive datasets, and with KEPT, the top study in the literature, creating a fair comparison environment.

5. Results

The results of each model were assessed using metrics such as BCE with logit loss, accuracy, recall, precision, and micro F1 score. The micro F1 score is a variant of the F1 score that takes into account the overall performance of a model across all classes in a multi-class classification task. It is calculated by aggregating the true positives, false positives, and false negatives over all classes before computing the F1 score. It is useful when there is a class imbalance in the dataset because it equally weighs the contribution of each class to the overall F1 score, regardless of their representation in the data.

Table 4
Results showing the advantage of using both data types per each model. For each metric, the best-performing model is highlighted with bold typeface.

			Loss (BCE with Logit)	Accuracy	Precision	Recall	Micro F1
Data Type	Dataset	Model					
Only Numerical Data	Dataset A	GRU	13.45	0.89	0.36	0.61	0.45
Only Discharge Summaries	Dataset A	Clinical BERT	9.47	0.93	0.44	0.69	0.54
	Dataset A	Clinical Longformer	10.61	0.92	0.56	0.75	0.64
	Dataset C	KEPT	9.16	0.93	0.74	0.72	0.73
Numerical Data & Discharge Summaries	Dataset A	Two-stream Model (Clinical BERT)	9.66	0.93	0.53	0.71	0.60
	Dataset A	Two-stream Model (Clinical Longformer)	9.32	0.93	0.61	**0.76**	0.69
	Dataset C	Two-stream Model (KEPT)	**8.67**	**0.94**	**0.76**	0.74	**0.75**

Table 5
Comparison of Results with Other Works in Literature.

Work	Micro F1 Score	Data Type	Dataset
This Work (KEPT & Numerical Classifier)	**0.751**	**Text & Numerical**	**Dataset C**
KEPT	0.733	Text	Dataset C
Zheng Yuan [6]	0.725	Text	Dataset B
Thanh Vu [5]	0.716	Text	Dataset B
Yang Liu [23]	0.717	Text	Dataset B
This Work (Clinical Longformer & Numerical Classifier)	**0.691**	**Text & Numerical**	**Dataset A**
Xiancheng Xie [3]	0.684	Text	Dataset B
Fei Li [27]	0.670	Text	Dataset B
Clinical Longformer	0.640	Text	Dataset A
James Mullenbach [4]	0.633	Text	Dataset B

For each model, the state that yielded the highest validation micro F1 score was chosen as the final model. Table 4 presents the outcomes obtained through these methods. Upon analyzing the results, numerical data has high accuracy but low precision values, resulting in low F1 scores. Therefore, it becomes evident that using numerical data alone is insufficient for the estimation task. However, when these data are combined with text-based data, a significant overall improvement can be observed. The micro F1 score sees a remarkable boost of nearly 5%. Thus, it can be asserted that the targeted outcome has been achieved, and the top 50 ICD-9 prediction task has yielded the best result in terms of micro F1 score.

Further, Table 5 provides valuable information by comparing the results of different studies in the literature with the results of the current study. The data in the table shows the performance of various models used in different studies in terms of the micro F1 Score, which gives the overall performance of multi-class classification models. Looking at the table, it is evident that the performance of the two different models used in this study is notably better compared to other studies in the literature. Our work with KEPT and the numerical classifier achieved a micro F1 Score of 0.751 by using both the text and numerical data types. Similarly, our other model using the Clinical Longformer and the numerical classifier obtained a micro F1 Score of 0.691 using the same data type. The comparison table also presents the performance of models used in other studies in the literature. Most of these models rely solely on text data, and their micro F1 Scores range from 0.633 to 0.733. These results suggest that an approach incorporating both text and numerical data is more effective in achieving better outcomes.

In this study, it was also revealed that the maximum token length produced by NLP algorithms is also important. The fact that the patient discharge notes were too long to be represented by 512 tokens caused the result from the Clinical BERT method to be more unsuccessful. In addition, although the maximum number of tokens was increased to 1024 with two Clinical BERT models for vectorization of discharge summaries, the results were insufficient. This inference is also similar to the conclusion in Yang et al.'s [7] study. Moreover, Biswas et al. [28], Zhang et al. [29], Pascual et al. [30] also showed the necessity of using higher dimensional token vectors.

It is also important to examine the results in more detail for each ICD Code. As an example, the resulting confusion matrices for the numeric-based GRU, text-based Clinical Longformer, and ensemble model are reported in Figs. 4, 5 and 6 respectively. These matrices show both the actual and expected results for every disease. Predicted outcomes are shown in columns, and actual labels are shown in rows. These matrices contain actual and predicted outcomes for each disease. Columns represent predicted results, rows represent actual labels. For each ICD code, the top left cell represents true negatives, which are the cases in which the model accurately predicted that the ICD code would not be present. The top right cell represents false positives, indicating the instances where the model incorrectly predicted the presence of the ICD code. The bottom left cell represents false negatives, showing the instances where the model failed to predict the ICD code when it was actually present. The bottom right cell represents true positives, which means the instances where the model correctly predicted the presence of the ICD code. When these matrices are examined, although numerical data-based methods on a class basis are not very successful, if these data are used with the help of the ensemble method, there has been a noticeable improvement in many classes. At this point, the most improving classes are Thrombocytopenia (2875) and Hyposmolality (2761). The number of true positives for thrombocytopenia was 17 in the numerical data-based model and only

Fig. 4. Confusion Matrix for Numerical Data Based GRU Model.

Fig. 5. Confusion Matrix for Text Data Based Clinical Longformer Model.

Fig. 6. Confusion Matrix for Both Numerical & Text Data Based Ensemble Model.

1 in the text-based model, while the ensemble model reached 18 true positives. Roughly, it can be said that some patients who cannot be identified from text data can be detected when numerical data are added to the ensemble model. Also, in addition to this improvement, almost no loss in true negative values was observed. Similar results can be obtained from confusion matrices for Hyposmolality and some other diseases. An important inference can be made as follows: text-based data and numeric data have a structure that compensates for each other's deficiencies in ICD estimation, instead of giving results that repeat each other.

6. Discussion

ICD code prediction is a complex task due to the vast number of classes, but it can greatly benefit doctors by automating the process. The advantages of automatic ICD code prediction include disease detection, providing suggestions to doctors when entering their own procedure's ICD codes, and giving reminders to healthcare professionals regarding which tests to administer based on tests given to similar patients. However, this task is challenging due to the extensive number of ICD classes, the vast array of tests conducted at hospitals with missing data where not all tests are available for all patients, and the possibility of multiple diseases within a single patient.

This study introduces a novel two-stream method for ICD-9 code prediction, but there are notable limitations: reliance on the specific MIMIC-III dataset may limit generalizability, the model's performance could vary in more complex medical scenarios, and the method demands substantial computational resources. While a significant advancement, these limitations emphasize the need for future research to enhance its applicability and accuracy in diverse healthcare settings.

7. Conclusion

In this study, we analyzed a substantial dataset of adults in an intensive care unit and proposed preprocessing methods, as well as natural language and machine learning models, to predict the ICD-9 codes of patients based on this data. The primary objective of this research is to enhance the performance of NLP methods using text-based data by incorporating numerical data in a two-stream network. To achieve this, we proposed both text and numerical-based methods and combined them to create a model that achieved the best results. For the two-stream models, a ratio of 0.75 with GRU and KEPT-based models yielded the best micro F1 score.

As a result, our study provides an effective and comprehensive approach for achieving improved results in the multi-class classification task of ICD code prediction. This research holds significant potential in assisting medical professionals with accurate ICD code predictions, ultimately leading to better patient care and outcomes; however, its potential should still be investigated across other ICU units and even other countries.

CRediT authorship contribution statement

Mustafa Arda Ayden: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Mehmet Eren Yuksel:** Writing – review & editing, Supervision, Project administration, Investigation, Conceptualization. **Seniha Esen Yuksel Erdem:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Assoc. Prof. Seniha Esen Yüksel Erdem would like to acknowledge the support of the Science Academy, Türkiye under the BAGEP Award 2023.

References

[1] N. Nath, S.-H. Lee, I. Lee, Application of specialized word embeddings and named entity and attribute recognition to the problem of unsupervised automated clinical coding, Comput. Biol. Med. 165 (2023) 107422.
[2] A.E. Johnson, T.J. Pollard, L. Shen, L. wei, H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (May 2016).
[3] P. Xie, E. Xing, A neural architecture for automated ICD coding, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018.
[4] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1101–1111.
[5] T. Vu, D.Q. Nguyen, A. Nguyen, A label attention model for icd coding from clinical text, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, 2020, pp. 3335–3341.
[6] Z. Yuan, C. Tan, S. Huang, Code synonyms do matter: multiple synonyms matching network for automatic ICD coding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 808–814.
[7] Z. Yang, S. Wang, B. Rawat, A. Mitra, H. Yu, Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing 2022, 2022, pp. 1767–1781.

[8] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, J. Biomed. Inform. 83 (2018) 112–134.
[9] J.R.L. Gall, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, JAMA J. Am. Med. Assoc. 270 (24) (1993) 2957–2963.
[10] J.L. Vincent, R. Moreno, J. Takala, S. Willatts, A.D. Mendonça, H. Bruining, C.K. Reinhart, P.M. Suter, L.G. Thijs, The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, Intensive Care Med. 22 (7) (1996) 707–710.
[11] H.M. Marsh, I. Krishan, J.M. Naessens, R.A. Strickland, D.R. Gracey, M.E. Campion, F.T. Nobrega, P.A. Southorn, J.C. McMichan, M.P. Kelly, Assessment of prediction of mortality by using the apache ii scoring system in intensive-care units, Mayo Clin. Proc. 65 (12) (1990) 1549–1557.
[12] E. Hossain, R. Rana, N. Higgins, J. Soar, P.D. Barua, A.R. Pisani, K. Turner, Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review, Comput. Biol. Med. 155 (2023) 106649.
[13] S. Ji, M. Hölttä, P. Marttinen, Does the magic of bert apply to medical code assignment? A quantitative study, Comput. Biol. Med. 139 (2021) 104998.
[14] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78.
[15] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.
[16] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 1, Association for Computational Linguistics, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[17] M. Jin, M.T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, T. Doman, A. Ravi, M. Liger, T. Kass-hout, Improving hospital mortality prediction with medical named entities and multimodal learning, 2018.
[18] M. Chen, Efficient vector representation for documents through corruption, in: 5th International Conference on Learning Representations, 2017.
[19] Y. Shen, H. Yun, Z. Lipton, Y. Kronrod, A. Anandkumar, Deep active learning for named entity recognition, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 252–256.
[20] Z. Lin, M. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, Toulon, France, 2017.
[21] A.D. Reys, D. Silva, D. Severo, S. Pedro, M.M. de Sousa e Sá, G.A.C. Salgado, Predicting multiple icd-10 codes from Brazilian-Portuguese clinical notes, in: R. Cerri, R.C. Prati (Eds.), Intelligent Systems, Springer International Publishing, Cham, 2020, pp. 566–580.
[22] L. Franz, Y. Shrestha, B. Paudel, A deep learning pipeline for patient diagnosis prediction using electronic health records, 2020.
[23] Y. Liu, H. Cheng, R. Klopfer, M.R. Gormley, T. Schaaf, Effective convolutional attention network for multi-label clinical document classification, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5941–5953.
[24] Y. Li, R.M. Wehbe, F.S. Ahmad, H. Wang, Y. Luo, Clinical-longformer and clinical-bigbird: transformers for long clinical sequences, arXiv preprint, arXiv: 2201.11838, 2022.
[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, 2013.
[26] I. Beltagy, M.E. Peters, A. Cohan, Longformer: the long-document transformer, arXiv:2004.05150, 2020.
[27] F. Li, H. Yu, Icd coding from clinical text using multi-filter residual convolutional neural network, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.
[28] B. Biswas, T.-H. Pham, P. Zhang, Transicd: transformer based code-wise attention model for explainable icd coding, in: A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, D. Riaño (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham, 2021, pp. 469–478.
[29] Z. Zhang, J. Liu, N. Razavian, BERT-XML: large scale automated ICD coding using BERT pretraining, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 24–34.
[30] D. Pascual, S. Luck, R. Wattenhofer, Towards BERT-based automatic ICD coding: limitations and opportunities, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 54–63.