Neural Language Modeling of Unstructured Clinical Notes for Automated Patient Phenotyping

Akshara Prabhakar*, Shidharth S*, Sowmya Kamath S
Healthcare Analytics & Language Engineering (HALE) Lab, Department of Information Technology,
National Institute of Technology Karnataka, Surathkal, Mangalore 575025, INDIA
Email: aksharap.181it132@nitk.edu.in, shidharth.181it243@nitk.edu.in, sowmyakamath@nitk.edu.in

***Abstract*—The availability of huge volume and variety of healthcare data provides a wide scope for designing cutting-edge clinical decision support systems (CDSS) that can improve the quality of patient care. Identifying patients suffering from certain conditions/symptoms, commonly referred to as phenotyping, is a fundamental problem that can be addressed using the rich health-related data collected for generation of Electronic Health Records (EHRs). Phenotyping forms the foundation for translational research, effectiveness studies, and is used for analyzing population health using regularly collected EHR data. Also, determining if a patient has a particular medical condition is crucial for secondary analysis, such as in critical care situations to predict potential drug interactions and adverse events. In this paper, we consider all categories of unstructured clinical notes of patients, typically stored as part of EHRs in the raw form. The standard MIMIC-III dataset is considered for benchmark experiments for patient phenotyping. Experiments revealed that our proposed models outperformed state-of-the art works built on vanilla BERT & ClinicalBERT models on the patient cohort considered, measured in terms of standard multi-label classification metrics like AUROC score (improvement by 6%), F1-score (by 4%), and Hamming Loss (by 17%) when we considered only patient discharge summaries and radiology notes. Further experiments with other note categories showed that using discharge summaries and physician notes yields significant improvements on the entire dataset giving 0.8 AUROC score, 0.72 F1 score, 0.09 Hamming loss.**

***Index Terms*—clinical decision support systems, patient phenotyping, unstructured text modeling, healthcare analytics**

I. Introduction

Electronic Health Records (EHRs) are a set of clinical data related to an individual patient's medical history, storing vital information pertaining to patients' primary care and healing processes. It includes structured data (like demographics, ICD codes, lab test results) and unstructured data in the form of clinical notes. Unstructured data comprises a significant part of structured EHRs manually coded by trained medical record department staff, and provide a rich source of patient-specific data like doctors' notes, nurses' notes, radiology reports, discharge summaries etc. Leveraging such unstructured data for secondary uses such as medical diagnosis, ICU mortality prediction, and clinical decision making requires several processing steps to transform them to a usable form.

Intensive Care Units (ICUs) are limited-resource, expensive environments where agile and accurate decision-making is crucial. However, most decision-making is often made spontaneously in high uncertainty settings and based on the clinician's prior experience. In such a scenario, discovering patient phenotypes is very helpful to determine how individual patients would respond to certain drugs and how they might react to different interventions. A phenotypic abnormality in medical settings is a deviation from normal human physiology, morphology, or behavior [1] and accurate phenotyping is a significant component of a doctor's daily job. Effective utilization of the widely adopted EHRs could aid in phenotyping.

Early EHR-based phenotyping systems primarily utilized rule-based techniques for patient categorization [2]. Most existing works use patient-specific information available in structured sources only, however, ignoring the rich patient data latent in unstructured sources can result in loss of vital information, affecting the creation of accurate classification systems. Recently, many approaches adopted Natural Language Processing to extract features from narrative text and for utilizing them for phenotyping. The free text's complex structural and temporal nature makes modeling the latent knowledge in patients' unstructured case reports challenging. Moreover, this is compounded by the fact that EHRs contain many distinct combinations of phenotypes, and their distribution across various data sources is skewed, with many labels occurring in a single text note.

Building upon existing works, we adopt and extend various pre-trained models like BERT and models trained on the clinical and biomedical domain such as ClinicalBERT, BioBERT, and BlueBERT for phenotyping patients based on the unstructured data obtained from the standard MIMIC-III dataset. The models are trained to learn from manually assigned phenotype annotations to attend to medical terms and exploit patient information from various note categories. Our work focuses on design of models for phenotype prediction, exploiting various BERT based models to perform extensive experimentation on the MIMIC-III dataset. We leverage different types of clinical note categories and Skip-Convolution with different attention mechanisms to effectively extract information from the notes.

The remainder of this article is organized as follows. Section II presents a discussion of relevant existing works in the field of interest. In Section III, we describe the dataset specifics in detail and also elaborate on the proposed methodology for language modeling of unstructured clinical notes for enabling automated phenotype prediction. Section IV presents the de-

*Equal contribution

tails of experimental evaluations performed and the observed results, followed by conclusions and directions for future work.

II. Related Work

For the task of automated patient phenotyping, the earliest works relied on rule-based mechanisms for the prediction. The most notable work is by Nguyen et al. [3], who proposed a rule-based classification system to identify various stages of lung cancer using text notes. They utilized MEDTEX, a tool consisting of modules for mapping text to the SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) [4] taxonomy terms. Harini et al. [5] used autoencoders for physiological time series signal reconstruction, which compress the inputs into a lower dimensional embedding. These low-dimensional embeddings are then passed through LSTM to get predictions. Rahman et al. [6] used prior medical knowledge from patients previously admitted, based on similarity for phenotyping. Here, similarity is determined by using the adjacency matrix of a created nearest neighbor graph to identify similar groups of ICU patients.

Gehrmann et al. [7] utilized CNNs for ten phenotyping tasks using data from 1,610 discharge summary notes, and compared them with concept extraction methods. Some works [8], [9] exploit noisy data from both structured and notes to learn from large amounts of imperfect data for phenotyping. Liu et al. [10] presented a binary classification problem for predicting readmission of heart failure patients from discharge summary notes. They used a Word2vec model fine-tuned on the PubMed dataset. Chen et al. [11] used EHR records to predict phenotypes, using Latent Dirichlet Allocation (LDA). Liu et al. [12] also used EHR to predict phenotypes, and introduced temporal graphs to better capture the longitudinal and heterogeneous properties of the EHR.

Closest to our work is the methodology adopted by Mulyar et al. [13], who used the BERT language model and experimented with mechanisms to summarize the entire clinical document. They considered the sequence of CLS tokens from various fragments, such as mean, concatenation and passed these through a LSTM layer, which improved performance on N2C2 dataset substantially. In contrast to their approach, we adopt transformer based architectures trained on unstructured clinical notes to generate clinical encodings for predicting phenotypes. We incorporate self and cross-attention mechanisms across different notes for emphasizing terms that are common and more indicative of a phenotype to enhance the prediction.

III. Materials & Methods

A. Data Preparation

For the experimental evaluation, we utilized the MIMIC-III [14] dataset, which contains clinical data relating to 61,532 critical care ICU patients. The *noteevents* table was utilized to obtain clinical notes of the patients, which contained various categories such as discharge summary, radiology, nursing and physician notes, corresponding to different `hadm_ids` (hospital admissions). We performed experiments considering the discharge summary independently, and then along with both discharge summary and radiology category (whenever present for a patient). For obtaining phenotype labels, we use the annotation provided by Moseley et al. [15]. There are 16 phenotype categories present in the manually annotated dataset of which 1 column is *Unsure* which we dropped due to lack of clarity. The remaining table is merged with above prepared notes, according to the hadm_ids. This complete dataset is denoted by D_{full} and similar to previous works, we consider the top-10 phenotype categories. Table I presents the statistics of phenotype categories distribution in the dataset. Most existing works utilize only a subset of the patients from the MIMIC-III dataset termed "frequent flyers" ($\geq$ 3 ICU visits in a year), as introduced by Gehrmann et al. [16]. It contains the discharge summary of 1,610 patients. We refer to this subset as D_{sub}.

TABLE I: Phenotype Distribution in D_{full}

Sl. No.	Phenotype Category	Frequency
1	CANCER	59
2	HEART DISEASES	145
3	LUNG DISEASES	83
4	ALCOHOL ABUSE	90
5	NEUROLOGICAL DISEASES	153
6	CHRONIC PAIN	126
7	DEPRESSION	195
8	OBESITY	42
9	OTHER SUBSTANCE ABUSE	67
10	PSYCHIATRIC DISORDERS	110

B. Preprocessing

Firstly, we performed standard preprocessing on the error-free notes (retaining notes having 0 in the *iserror* column of the *noteevents* table) to clean the text corpus: tokenization, removal of stopwords, stemming, and lemmatization. Initially, we converted the text to lowercase letters. We removed punctuation marks, special characters and numbers. Next, we removed stop words, as they would not help in extracting phenotype features. Then, all words are converted to their lemma, as lemmatization can help in reducing the complexity of the model. Since the notes contain a lot of medical jargon, which are not used in found in common English vocabularies, we trained a new vocabulary for the discharge summary notes using BertWordPieceTokenizer from Transformers library. For word embeddings, we used Xavier weights initialized embedding layer. This custom tokenizer was used to encode texts.

C. Models

The proposed model consists of 3 components: Encoder, Attention and the Prediction module. The architecture of the proposed model is shown in Fig. 1. We discuss each of the components in detail below.

1) Encoder Module.: To get the encodings, we try using the [CLS] token from only the last layer of the proposed model and also the mean pooling of [CLS] tokens from every intermediate layer, which performed the best during our experiments.

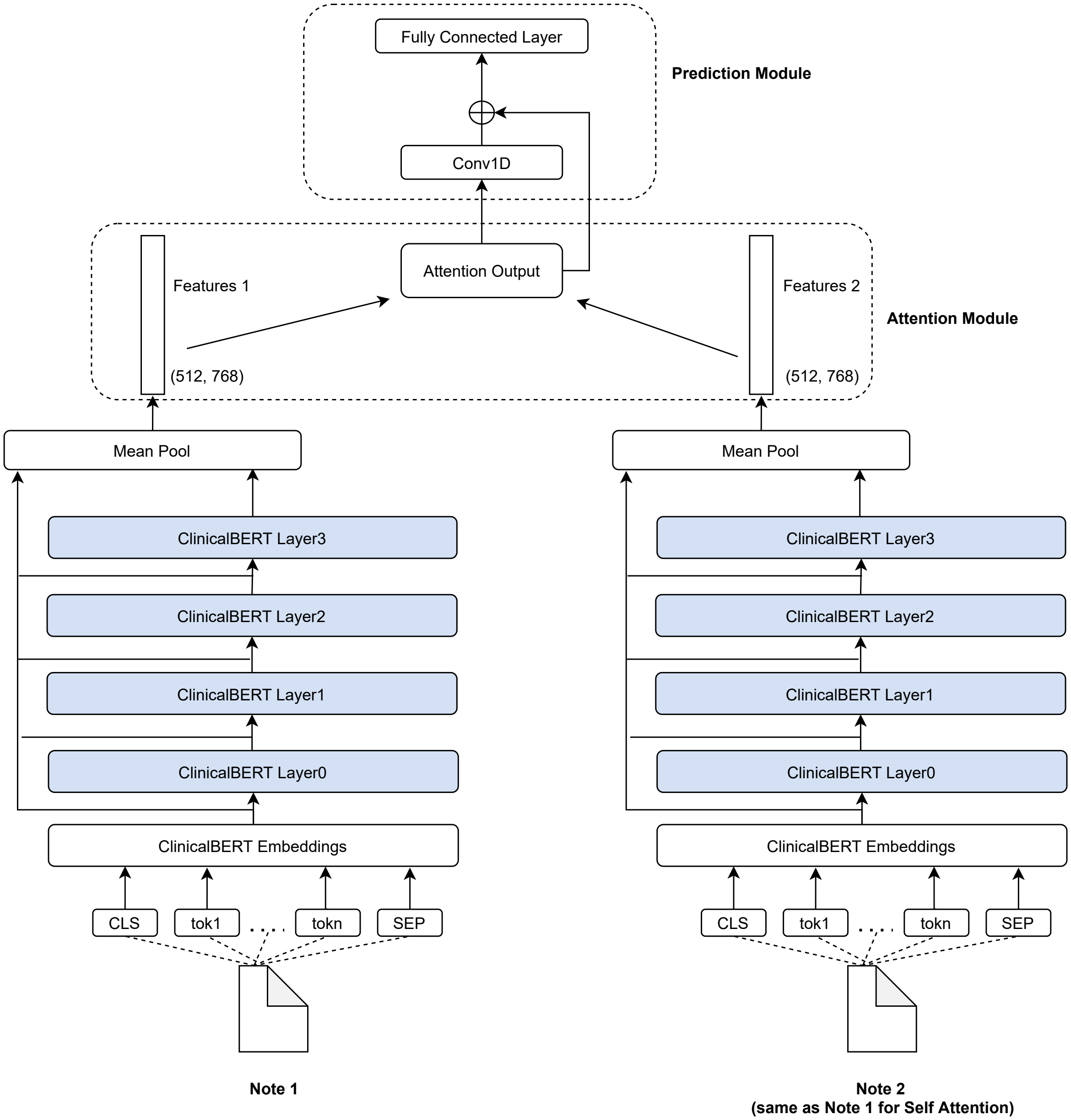

Fig. 1: Overview of the proposed model

(a) *BERT Encoder.* BERT (Bidirectional Encoder Representations from Transformers) [17] proposes a masked language modeling (MLM) objective, where some of the input sequence tokens are randomly masked, and the task is to correctly predict the masked indices taking the augmented input sequence. It is also trained on Next Sentence Prediction (NSP) where the model gets a sentence pair as input, with the training objective being predicting if a sentence can be the next subsequent sentence in a document.

(b) *ClinicalBERT Encoder.* We experimented with models like ClinicalBERT, Bio-BERT, PubMedBERT, Bionlp's BlueBERT, with the first 4 layers, as deeper architectures caused overfitting. We found that embeddings obtained from ClinicalBERT gave the best results. ClinicalBERT [18] is trained on all MIMIC-III notes. It has 12 layers, 768 hidden states and 12 heads, of which we use the first 4 layers, 4 heads and all the hidden states. We found this encoder to capture better clinical similarity compared to models such as FastText and Word2Vec. ClinicalBERT is an application of BERT in the medical domain and has been used on a wide variety of tasks such as Hospital readmission prediction, Diagnostic code prediction, Disease modelling etc.

2) *Cross & Self Attention Module:* Next, the encodings are passed to a multi-headed attention [19] layer with 4 heads. Both the discharge notes as well the radiology notes are passed to the encoder to get separate outputs. We give the output

from the discharge notes as Query Q and output from the radiology notes as Key K and Value V. Here, all vectors have the same dimension of 768. The output of the multi-headed (MH) attention layer so obtained is then averaged along the sequence length. This is done because having the key vectors as encoded radiology notes and the query vector as encoded discharge summary notes, the common medical terms are coded similarly and hence are given more importance while performing the attention mechanism. We refer to this form of attention as Cross-Attention. We also experimented with using only discharge summary notes, in which case it reduces to Self-Attention with K, Q, V all being the same.

$$MH(Q, K, V) = Concat(head_1, head_2, head_3, head_4)W^o \tag{1}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

3) Prediction Module: For all experiments, we used the attention outputs computed from above as inputs. For the simple fully connected layer, we simply passed the inputs via a linear layer having 10 classes. Next, we use a skip connection and add the attention outputs with the outputs from the encoder. We pass this to a Conv1D layer with same padding and a kernel size of 3. This is indicated as skip-Conv in tables II and III. Skip Connections [20] as shown in Fig. 2 was first introduced in the vision domain, however the concept can be extended to any domain. The performance of a model drops with the increase in depth of the architecture, known as the degradation problem. We use skip connections here to re-enforce the original discharge summary encodings in the final output. The idea behind this is to execute a weighted skip connection of sorts where information from the Conv1D layer is combined with the original features from the encoder thus retaining an essence of the original embeddings. This was then passed through an average pool layer. The output of this was finally passed through a linear layer from where predictions are generated.

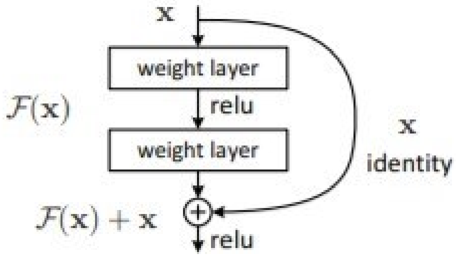
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2: Skip Connections [20]

On the final layer, we use a sigmoid activation function with threshold set at 0.5. The weighted Binary cross entropy loss function is used, and the weights are obtained by calculating the sum of all positive instances across all rows and columns and dividing the same by the the size of the matrix. The value computed indicated an imbalance in the classes, so to eliminate it, the weight of the class containing the negative instances was set to the weight of the positive instances (inverse of the negative instance weight [21]).

IV. Experimental Results & Analysis

We used Sklearn, Pandas, Gensim and PyTorch libraries for developing the proposed models. All experiments were performed on a server with 2 NVIDIA M40 GPUs. The optimizer used was *AdamW*, Cosine scheduler and loss function was *binary_crossentropy*. The vocabulary size is the same as that of the size of the custom tokenizer used, with a max sequence length of {512, 1000, 3000} and a batch size of 1 with learning rate set as 2e-5. The models were trained for 20-50 epochs with early stopping when there is no improvement in F1-score on the validation set over 3 consecutive epochs. Standard scoring metrics such as F1 score, Hamming loss and AUROC scores are used to assess the performance. We also experimented with various activation functions such as GELU, LeakyRelu, which however did not improve the performance. Table II reports the results obtained and the effect of the constituent modules on all the patients, i.e. D_{full}.

A. Baseline Models

To benchmark our approach, we compare some of our best performing models in Table II with the following baseline models on D_{sub}:

1) *CNN* [16]: A convolution neural network by Gehrmann et al. to identify patient phenotypes using discharge reports.
2) *ws-CNN* [22]: A CNN with three different filter sizes with a combination of word and sentence level embeddings by Yang et al. to identify patient phenotypes from discharge reports.
3) *ClinicalBERT* [18]: They use pretrained BERT on clinical notes and fine-tune the network for predicting hospital readmissions at various time points.
4) *EnCAML* [23]: proposed a multi-channel, variable-sized CNN based attention model to predict ICD-9 codes. It attends to clinical notes and extracts relevant snippets from them to map to the medical codes.
5) *ClinicalBERT based f_{mean}* [13]: They divide the entire clinical document into chunks and use various approaches to combine important information from them using the sequence of CLS tokens, with the best performance obtained when a mean is taken.

As can be observed from Table III, the proposed model outperforms existing state-of-the-art works on all scoring metrics, which underscores the effectiveness of our model. Our model leverages the attention on the encodings after they are generated as well. Apart from this we use a pre-trained model trained on a clinical corpus to get the initial word embeddings.

B. Discussion

From Table II, we observed improvement in results when both radiology and discharge summary notes were used together i.e. Cross Attention. The intuition behind this is that when both the notes using the encoded representation of discharge summary to query over the representation from the

TABLE II: Performance of various models on D_{full} for the Phenotyping task

No.	Model	F1 Score (avg)	Hamming Loss	AUROC (micro)
1	BERT base (discharge summary)	0.27	0.15	0.41
2	BERT + Self Attention	0.44	0.14	0.64
3	BERT + Cross Attention	0.53	0.11	0.69
4	ClinicalBERT + Self Attention	0.52	0.14	0.69
5	**ClinicalBERT + Cross Attention**	**0.66**	**0.09**	**0.77**
6	ClinicalBERT + Cross Attention + skip-Conv	0.63	0.10	0.76
7	ClinicalBERT + Self Attention + skip-Conv	0.56	0.12	0.71

TABLE III: Comparison with baseline models on D_{sub} for the Phenotyping task

No.	Model	F1 Score (avg)	AUROC (micro)	Hamming Loss
1	CNN based [16]	0.59	-*	-*
2	ws-CNN [22]	0.56	-*	-*
3	ClinicalBERT	0.46	0.61	0.19
4	EnCAML [23]	0.15	0.54	0.14
5	f_{mean} [13]	0.53	0.71	0.12
6	**ClinicalBERT + Self Attention + skip-Conv (ours)**	**0.61**	**0.75**	**0.10**
7	ClinicalBERT + Cross Attention + skip-Conv (ours)	0.55	0.70	0.11

**value not reported.*

radiology notes are used, the fragments in both notes which are common that indicate the presence of some target disease gets more attention. Due to this, a richer set of features are generated which help in classification of the target phenotypes. We also observed that the hamming loss improved when class weights [21] are used to handle the imbalance of phenotype categories, thus predicting the phenotypes more accurately.

We observed some interesting trends in the performance of our models on the entire D_{full} and the subset D_{sub}. For instance, the Cross Attention works better in the case of the full dataset, probably because of overlapping information in both notes present of the patients in the subset, as they visited the ICU more than 3 times. We also a notice a drop in the performance of the Skip-Conv module when the full dataset is used, however, the model performed well in the case of the subset of patients.

We further experimented with different types of notes apart from discharge summary, like radiology, physician and nursing in the Cross Attention module, to study the effect of using different notes, the results of which are tabulated in Table IV. All these experiments are carried out on the model that achieved best results (Ref: Table II). In each scenario, we consider the patients who have both types of note categories available, thus resulting in differing data sizes for each experiment.

TABLE IV: Experiments with different note categories

Query	Key	F1 Score (avg)	Hamming Loss	AUROC (micro)
Radiology	Physician	0.24	0.16	0.56
Physician	Nursing	0.32	0.15	0.60
Physician	Discharge	0.44	0.14	0.64
Nursing	Discharge	0.65	0.11	0.75
Discharge	Nursing	0.43	0.15	0.64
Physician	Radiology	0.13	0.18	0.52
Nursing	Physician	0.43	0.15	0.64
Discharge	**Physician**	**0.72**	**0.09**	**0.80**
Discharge	Radiology	0.66	0.09	0.77
Radiology	Discharge	0.21	0.16	0.55

From this, we observe that interchanging the Key (K) and Query (Q) vectors among different notes has varying effects. Mathematically, this is because the final context vector (the weighted average of the Value vector V passed to the Prediction module) changes when this interchange is made (refer Eq. 1). For instance, consider the category pair *Discharge* and *Physician*. Discharge summary is written for a patient towards the end of their hospital admission and is concise and concrete. On the other hand, physician notes include the initial findings upon examination which are quite broad. Using Discharge as Q to attend over Physician notes K, gives more focus on the final findings and takes further context from K to enhance the predictions. This is also justified by the fact using Discharge summary notes alone, with self attention results in a lower performance. If we switch and take Physician as Q, it would give lesser emphasis to the terminal findings which are more indicative of the actual patient condition, leading to a drop in the scores. Another way to explain this is by drawing a parallel to translation tasks, where, query, key and values are used in a similar way. The multiheaded attention module here essentially maps queries against a set of keys. In certain cases, some mappings cannot take place due to absent information in the keys. This argument is also further strengthened by the fact that, during translation and back-translation, the original sentence and the newly generated sentence from back-translation don't match even though they may essentially mean the same.

We also found that using self attention on notes other than discharge summary gave poor results. This leads us to believe that the notes taken farthest down the timeline since the patient was admitted might have more accurate information on their condition. If we consider a timeline of a patients stay in the hospital then discharge summary notes are prepared at the time of discharge, hence this is last in the timeline, whereas, nursing notes or physician notes are created during the admission period, thus marking the beginning of our timeline. The radiology notes are slotted towards the middle of the timeline as it is done after consulting with a doctor and hence after Physician notes.

V. Conclusion & Future Work

In this work, several methods and approaches for unstructured clinical text based neural modeling for the patient phentoyping task was explored, through the use of transformer based architectures. Our approach is built on a pre-trained model trained on clinical notes to obtain clinical encodings. These encodings are passed through a self/cross-attention layer and the outputs of this attention layer are combined with the encodings generated before to reinforce past learning. This enables the terms that are unique to better indicate the prediction of a relevant phenotype. The proposed model outperformed the state-of-the-art models on the multilabel classification task of phenotyping. Since ICD-9 coding is also a multi-label prediction task, this approach could be extended for that as well. An interesting direction of future research could be trying to model the problem by utilizing additional information from the structured data such as ICD-9 codes to enhance the performance.

References

[1] P. Robinson, "Deep phenotyping for precision medicine," *Human mutation*, vol. 33, pp. 777–80, 05 2012.

[2] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–230, 2014.

[3] A. N. Nguyen, M. J. Lawley *et al.*, "Symbolic rule-based classification of lung cancer stages from free-text pathology reports," *Journal of the American Medical Informatics Association*, vol. 17, no. 4, pp. 440–445, 07 2010.

[4] K. Donnelly *et al.*, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[5] H. Suresh, P. Szolovits, and M. Ghassemi, "The use of autoencoders for discovering patient phenotypes," 2017.

[6] A. Rahman, Y. Chang, B. Conroy, and M. Xu-Wilson, "Phenotyping with prior knowledge using patient similarity," in *Proceedings of the 5th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 126. PMLR, 2020, pp. 331–351.

[7] S. e. a. Gehrmann, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," 2018.

[8] V. Agarwal, P. Lependu, T. Podchiyska, R. Barber, M. Boland, G. Hripcsak, and N. Shah, "Using narratives as a source to automatically learn phenotype models," in *Workshop on Data Mining for Medical Informatics*, 2014.

[9] Y. Halpern, Y. Choi, S. Horng, and D. A. Sontag, "Using anchors to estimate clinical state without labeled data," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2014, pp. 606–15, 2014.

[10] X. Liu, Y. Chen, J. Bae, H. Li, J. Johnston, and T. Sanger, "Predicting heart failure readmission from clinical notes using deep learning," 2019.

[11] Y. Chen, J. Ghosh *et al.*, "Building bridges across electronic health record systems through inferred phenotypic topics," *Journal of Biomedical Informatics*, vol. 55, pp. 82–93, 2015.

[12] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. Association for Computing Machinery, 2015, p. 705–714.

[13] A. Mulyar, E. Schumacher, M. Rouhizadeh, and M. Dredze, "Phenotyping of clinical notes with improved document classification models using contextualized neural language models," *ArXiv*, vol. abs/1910.13664, 2019.

[14] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[15] W. J. Moseley E, Celi L A and D. F., "Phenotype annotations for patient notes in the mimic-iii database (version 1.20.03)," 2020. [Online]. Available: https://doi.org/10.13026/txmt-8m40.

[16] S. Gehrmann, F. Dernoncourt *et al.*, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLOS ONE*, vol. 13, pp. 1–19, 02 2018.

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

[18] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," 2020.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[21] A. P, S. S, G. S. Krishnan, and S. K. S, "Integrating structured and unstructured patient data for icd9 disease code group prediction," in *8th ACM IKDD CODS and 26th COMAD*. Association for Computing Machinery, 2021, p. 436.

[22] Z. Yang, M. Dehmer, O. P. Yli-Harja, and F. Emmert-Streib, "Combining deep learning with token selection for patient phenotyping from electronic health records," *Scientific Reports*, vol. 10, 2020.

[23] V. Mayya, S. K. S., G. S. Krishnan, and T. Gangavarapu, "Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries," *Future Generation Computer Systems*, vol. 118, pp. 374–391, 2021.