Transformers for Multi-label Classification of Medical Text: An Empirical Comparison

Vithya Yogarajan[0000−0002−6054−9543], Jacob Montiel[0000−0003−2245−0718], Tony Smith[0000−0003−0403−7073], and Bernhard Pfahringer[0000−0002−3732−5787]

Department of Computer Science, University of Waikato, New Zealand
vy1@students.waikato.ac.nz

Abstract. Recent advancements in machine learning-based multi-label medical text classification techniques have been used to help enhance healthcare and aid better patient care. This research is motivated by transformers' success in natural language processing tasks, and the opportunity to further improve performance for medical-domain specific tasks by exploiting models pre-trained on health data. We consider transfer learning involving fine-tuning of pre-trained models for predicting medical codes, formulated as a multi-label problem. We find that domain-specific transformers outperform state-of-the-art results for multi-label problems with the number of labels ranging from 18 to 158, for a fixed sequence length. Additionally, we find that, for longer documents and/or number of labels greater than 300, traditional neural networks still have an edge over transformers. These findings are obtained by performing extensive experiments on the semi-structured eICU data and the free-form MIMIC III data, and applying various transformers including BERT, RoBERTa, and Longformer variations. The electronic health record data used in this research exhibits a high level of label imbalance. Considering individual label accuracy, we find that for eICU data medical-domain specific RoBERTa models achieve improvements for more frequent labels. For infrequent labels, in both datasets, traditional neural networks still perform better.

Keywords: Multi-label · Fine-tuning · Medical text · Transformers · Neural Networks

1 Introduction

There has been a significant advance in natural language processing (NLP) in the last couple of years. Transformers such as BERT models (Bidirectional Encoder Representations from Transformers) have outperformed state-of-the-art (SOTA) results [6, 7, 4]. Such advancements are not restricted to general-domain tasks. Biomedical and health-related domains have also seen evidence of improvements in some medical domain-specific tasks such as question answering and recognizing question entailment [3, 2, 9]. This research sets out to fill the gap in the use of transformers in multi-label medical domain-specific tasks for highly imbalanced datasets.

Multi-label problems predict multiple output variables for each instance. Consider a dataset $D = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$ with N samples, where $x^{(i)} = (x_1^{(i)}, ..., x_m^{(i)})$ and $y^{(i)} = (y_1^{(i)}, ..., y_l^{(i)})$. Each instance is associated with L labels, and each label is binary where $y_j^{(i)} \in 0, 1$. For example, given a patient admitted in a hospital with chest pain, any other medical condition that the patient has, such as cholesterol, blood pressure, or obesity, can be considered as labels.

This research focuses on electronic health records (EHR) from two distinctly different large publicly available medical databases: MIMIC-III contains huge documents in a free-form medical text; eICU has concise, compressed medical data presented in the semi-structured form. Automatically predicting medical codes is the down-stream task for this research where we fine-tune pre-trained transformer models, and we present results for multi-label medical code classifications with the number of labels being 18, 93, 158, 316, and 923.

The contributions of this work are: (i) we analyse the effectiveness of using transformers for the task of automatically predicting medical codes from EHRs for multiple document lengths and number of labels; (ii) we demonstrate that for documents with sequence length truncated at 512 tokens, medical domain-specific transformer models outperform SOTA methods for multi-label problems with 18, 93 and 158 labels for both datasets; (iii) it is shown that for longer documents, larger multi-label problems, and infrequent labels, transformer models' F1 scores are not as good as the traditional word-embeddings-based SOTA neural networks.

2 Related Work

This research is motivated by the recent advancements of transformer models which have shown substantial improvements in many NLP tasks, including BioNLP tasks. With minimum effort, transfer learning of pre-trained models by fine-tuning on down-stream supervised tasks achieves very good results [3, 2]. For example, PubMedBERT [9] achieves SOTA performance on many biomedical natural language processing tasks such as named entity recognition, question answering and relation extraction and holds the top score on the Biomedical Language Understanding and Reasoning Benchmark (BLURB) [9].

Automatically predicting medical codes from EHRs has been studied over the years, where rule-based, machine learning-based and deep learning approaches have been proposed. Techniques including CNNs, RNNs and Hierarchical Attention Networks are some examples of deep learning approaches [16, 2]. Mullenbach et al. (2018) [17] present Convolutional Attention for Multilabel classification (CAML) which uses the MIMIC III dataset for ICD-9 code predictions. As mentioned by the survey of deep learning methods for ICD coding of medical documents presented by Moons et al. (2020) [16] CAML is considered the SOTA method for automatically predicting medical codes from EHRs.

There is some evidence of the use of transformer models in automatically predicting medical codes such as submissions to CLEF eHealth 2019 ICD-10 predictions from German documents [2, 19], and BERT and XLNet performance

on most frequent ICD-9 codes from MIMIC III with a maximum number of tokens set at 512 [20]. However, it is unclear how well transformer models can perform with long clinical documents and in multi-label problems with a large number of labels [20]. Also, many studies [20, 3] focus on high-frequency labels. Nonetheless, datasets such as MIMIC III and eICU consist of many infrequent labels where most codes only occur in a minimal number of clinical documents. This research presents results of multiple transformer methods and compares it with SOTA methods for various token lengths and number of labels.

For both word embeddings based networks and transformers, there is evidence to show domain-specific pre-trained models outperform general text pre-trained models [10, 9, 22]. This research uses word embeddings pre-trained on health-related text and transformers pre-trained on general and health-related data.

3 Data

Medical Information Mart for Intensive Care (MIMIC-III) [11, 8] is a publicly available large database from the MIT with de-identified medical text data of more than 50,000 patients. We make use of free-form medical text from the discharge summaries. Figure 1 (top) presents a small sample of a discharge summary. MIMIC III discharge summary length varies between 50 to 8500 tokens with an average pre-processed text length of 1500 tokens. There are approximately 9000 unique ICD-9 codes associated with the hospital admissions in this database, with more than one code assigned to each patient.

Electronic Intensive Care Unit (eICU) is a database formed from the Philips eICU program [8, 18], and contains de-identified data for more than 200,000 patients admitted to ICU. eICU data is found in tabular format with a drop-down menu. Sample text data is presented in Figure 1 (bottom). The length of medical text ranges from 10 to 1350, with an average of 130 tokens. eICU contains 883 unique ICD-9 codes.

The frequency of ICD-9 codes in both MIMIC III and eICU is unevenly spread with a large proportion of the codes occurring infrequently. For example,

MIMIC III - Discharge Summary (sample text)
82 yo M with h/o CHF, COPD on 5 L oxygen at baseline, tracheobronchomalacia s/p stent, preseents with acute dyspnea over several days, and lethargy. This morning patient developed an acute worsening in dyspnea, and called EMS. EMS found patient tachypnic at saturating 90% on 5L. Patient was noted to be tripoding. He was given a nebulizer and brought to the ER.

eICU - Drop down menu (sample text)
Admission |Non-operative |Diagnosis |Cardiovascular |Sepsis, pulmonary |Non-operative Organ Systems |Was the patient admitted from the O.R. or went to the O.R. within 4 hours of admission? |No

Fig. 1: Sample data of MIMIC III (top) and eICU (bottom) obtained from the database. It includes acronyms and typos that are present in the data.

in MIMIC III and eICU only 0.02% and 0.2% of the codes are associated with at least 500 (1%) of the hospital admissions. One of the main reasons for the infrequent nature of medical codes in MIMIC III and eICU is because data are obtained from patients admitted in critical care. For this research, we consider each level of the ICD-9 hierarchy, as categorised by the World Health Organisation, as an individual flat multi-label problem. We remove all codes that occur in less than 10 unique hospital admissions. Consequently, our MIMIC III and eICU datasets contain 18 labels at level 1, 158 and 93 labels respectively at level 2, and 923 and 316 labels respectively at level 3.

4 Neural Network Algorithms

4.1 Transformers

Transformers [21] are one of the main recent developments in NLP which have achieved SOTA results in many language tasks [6, 9, 7]. Transformers are sequence-to-sequence models based on a self-attention mechanism. Given the linear projections Q, K, V, self-attention is computed as following [21]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where the input queries and keys are of dimension d_k, and values of dimension d_v. See Vaswani et al. (2017) [21] for details of the transformer architecture.

BERT [7] is a deep neural network model that applies bidirectional training of the transformer encoder architecture [21] to language modelling. The BERT model relies on two pre-training tasks, masked language modelling and next sentence prediction. The 12-layer BERT-base model with a hidden size of 768, 12 self-attention heads, 110M parameter neural network architecture, was pre-trained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia.

ClinicalBERT model follows the same model architecture as the BERT-base model and was continually pre-trained on all notes from MIMIC III [1] from the BERT weights. PubMedBERT [9] uses the same architecture as the BERT-base model. However, unlike ClinicalBERT, PubMedBERT is domain-specifically pre-trained from scratch using abstracts from PubMed and full-text articles from PubMedCentral to enable better capturing of the biomedical language [9].

RoBERTa [14] is a robustly optimized BERT approach with improved training methodology and 160GB of general-domain training data in comparison to the 16GB data used in BERT. BioMed-RoBERTa-base [10] is based on the RoBERTa-base [14] architecture. RoBERTa-base was continuously pre-trained using 2.68 million scientific papers from the Semantic Scholar corpus starting with the RoBERTa-base weights.

Longformer [4] is a transformer model that is designed to handle longer sequences without the limitation on the maximum token size of 512 set by other

transformers such as BERT. Longformer reduces the model complexity by reformulating the self-attention computation. This modified self-attention operation scales linearly with sequence length, instead of quadratically as in the original transformer models, making it possible to handle long documents. Longformer combines attention patterns such as sliding windows, dilated sliding windows and global attention (see Beltagy et al. (2020) [4] for more details). When compared to Equation 1, Longformer uses two sets of projections, one to compute attention scores for a sliding window and another for global attention, providing the needed flexibility for the best performance of downstream tasks [4]. Longformers can be used for other NLP tasks in addition to language models. When compared to Transformer-XL [6], which can also handle long documents, Longformer is not restricted to the left-to-right approach of processing the documents.

After pre-training the models, the transformers are fine-tuned on task-specific data. All the parameters are fine-tuned end-to-end. Pre-trained transformer models learn good, context-dependent ways of representing text sequences which can be used on a specific downstream task. The models only need to fine-tune their representations to perform a particular task. Compared to the pre-training cost of transformers, the subsequent fine-tuning is relatively inexpensive.

4.2 Traditional Neural Networks

TextCNN [12] combines a single layer of one-dimensional convolutions with a max-over-time pooling layer and one fully connected layer. If $x_{i:i+j}$ is a concatenation of words from a sentence, each word, x_i, x_{i+1}, ... is mapped to its embeddings using the lookup table of word embeddings. The final prediction is made by computing a weighted combination of the pooled values and applying a sigmoid function. In our experiments, we use TextCNN with four different window sizes where each window takes 2, 3, 4 or 5 words with 100 feature maps each; the drop out rate is set to 0.2 and the learning rate to 0.003.

Gated Recurrent Units (GRU) [5] are a type of recurrent neural networks, with fewer parameters in comparison to long short-term memory (LSTM) networks. Bidirectional GRU (BiGRU) considers sequences from left to right, and right to left simultaneously. The learning rate used for our experiments is 0.003.

Mullenbach et al. (2018) [17] present CAML which achieves SOTA results for predicting ICD-9 codes from MIMIC III data [16]. CAML combines convolution networks with an attention mechanism. A secondary module is used to learn embeddings of the descriptions of ICD-9 codes to improve predictions of less frequent labels and are used as target regularization. For each word in a given document, word embeddings are concatenated into a matrix and a one dimensional convolution layer is used to combine these adjacent embeddings. The document is represented by matrix $\mathbf{H} \in R^{d_c \times N}$ where d_c is the size of convolutional filter and N is the length of the document. Then a per-label attention mechanism is applied, where $\mathbf{H}^T \mathbf{u}_l$ is computed for a given label l and a vector parameter $\mathbf{u}_l \in R^{d_c}$. The resulting vector is passed through a softmax operation with an output α_l. The vector representation for each label is calculated using $\mathbf{v}_l = \sum_{n=1}^{N} \alpha_{l,n} h_n$. The probability for l is calculated using a linear layer and a

sigmoid transformation. A regularizing objective was added to the loss function of CAML with a trade-off hyperparameter. This variant is called Description Regularized-CAML (DR-CAML) [17]. The learning rate used for both CAML and DR-CAML in our experiments is 0.0001, and the regularization hyperparameter λ for DR-CAML is 0.01.

5 Experiments

We present results for multi-label medical code predictions for MIMIC III and eICU datasets. The number of labels being 18, 93, 158, 316, and 923. All experimental results presented are obtained from validations based on training-testing scheme, and are averaged over three runs. We explore a number of different transformer models and compare the performance to some traditional word embeddings based neural networks, including SOTA networks. The medical documents are truncated to a maximum number of tokens (512 and 4000). MIMIC III text was pre-processed by removing tokens that contain non alphabetic characters, including all special characters, and tokens that appear in fewer than three training documents. As eICU is already pre-processed extensively, no additional pre-processing was done for our research.

All neural network models presented in this research are implemented in PyTorch, and evaluations were done using sklearn metrics. All transformer implementations are based on the open-source PyTorch-transformer repository.[1] Transformer models are fine-tuned on all layers without freezing. As the optimizer we use Adam [13] with learning rates of 4e-6, or 4e-5. Training batch sizes were varied between 1 and 16, and the cut-off threshold was set to $t = 0.5$. Embeddings used for TextCNN, CAML, DR-CAML and BiGRU are health domain-specific fastText [15] pre-trained, skipgram word representation, 100-dimensional embeddings.

6 Results

Results for levels 1, 2 and 3 of the ICD-9 hierarchy, where each level is treated as an individual flat multi-label problem, for both eCIU and MIMIC III data are presented in Table 1. For eICU, we present results for 18, 93 and 316 labels. We find that using transformers for 18 and 93 labels, especially domain-specific models, result in performance improvements. We experimented with a maximum token length of 128, 512, and 1250 for eICU, and noticed a consistent improvement in performance between 128 and 512 tokens. However, there was no change between the micro and macro F1 scores for data truncated at 512 tokens and 1250 tokens. Due to space limitations, we only present results for the maximum token length of 512. It is important to notice that only 0.2% of the eICU data contains medical text with a sequence length greater than 512. This might explain the small variation in neural network performances when the maximum

[1] https://github.com/huggingface/transformers

Table 1: Micro and macro F1 scores for multi-label problem with labels ranging from 18 to 923 are presented for eICU (left) and MIMIC III (right) datasets. Bold is used to indicate the highest scores within the grouping of networks, and underline to indicate the best score across all presented. Reported results are from validations based on training-testing scheme, averaged over three runs.

	eICU - 93 Labels		MIMIC III - 158 labels			
	512 tokens		512 tokens		4000 tokens	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
TextCNN	0.54	0.30	0.62	**0.32**	0.69	0.39
CAML	**0.57**	0.31	**0.64**	**0.32**	**0.72**	**0.42**
DR-CAML	**0.57**	**0.32**	**0.64**	**0.32**	**0.72**	**0.42**
BiGRU	0.56	**0.32**	0.60	0.31	0.70	**0.42**
Longformer	**0.60**	0.28	0.64	0.35	**0.70**	**0.38**
BERT-base	0.59	0.28	0.62	0.37	n/a	n/a
ClinicalBERT	0.59	0.28	0.64	0.36	n/a	n/a
BioMed-RoBERTa-base	**0.60**	**0.32**	0.64	0.40	n/a	n/a
PubMedBERT	0.58	0.24	**0.65**	**0.41**	n/a	n/a

	eICU - 512 tokens				MIMIC III - 512 tokens			
	18 labels		316 labels		18 labels		923 labels	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
TextCNN	0.63	0.48	0.43	0.17	0.79	**0.70**	0.50	0.18
CAML	**0.65**	**0.51**	0.50	**0.20**	0.79	0.69	**0.54**	**0.19**
DR-CAML	**0.65**	**0.51**	**0.51**	**0.20**	**0.80**	**0.70**	0.53	**0.19**
BioMed-RoBERTa-base	**0.68**	**0.52**	**0.50**	0.13	0.79	0.72	0.52	0.15
Pub-MedBERT	**0.68**	**0.52**	**0.50**	**0.14**	**0.81**	**0.74**	**0.53**	**0.16**

sequence length is greater than 512 tokens. Compared to the word embeddings based methods, there is an improvement in micro-F1 when transformers are used. The overall best results are obtained using BioMed-RoBERTa-base for 93 labels, and Pub-MedBERT and BioMed-RoBERTa-base for 18 labels. However, for larger multi-label problem, such as the 316 labels, CAML and DR-CAML performs better with more significant differences in macro-F1 scores.

For MIMIC III, we present results for a maximum sequence length of 512 and 4000 tokens for 158 labels, and 512 tokens for 18 and 923 labels. As mentioned in Section 3, MIMIC III contains long documents and benefits from the increase in the length of maximum sequence size. Results using 4000 tokens are only presented for Longformer as the other transformer models are designed to handle a maximum of 512 tokens. Compared to the SOTA methods CAML and DR-CAML, most transformers show performance improvement for maximum sequence length of 512 tokens for 18 and 158 labels. For 158 labels macro-F1 of all transformers are considerably better than that of the SOTA methods, with PubMedBERT setting a new SOTA results for ICD-9 code prediction. Similarly, with 18 labels, PubMedBERT results are better than that of word embeddings-based methods for 512 tokens. However, as observed for eICU with 923 labels, none of the transformers perform as well as the traditional neural networks, when the number of labels increases. However, we have only explored a subset

of possible transformers. Future research might result in transformers that work well for multi-label problems with many infrequent labels.

Longformer is one of the very few transformers that can handle long documents. The model used in this research is pre-trained using general-domain data; however, like BERT and RoBERTa models, Longformer models trained on health domain-specific data may improve performance. To the best of our knowledge, there is no publicly available health domain-specific pre-trained Longformer, and it requires extensive resources to undertake such a task. Hence, we only present results for the general domain pre-trained publicly available model. It is essential to point out we also explored the option of using XLNet. However, a down-stream task for such large multi-label problem for text with tokens > 512 requires considerable computational power and time. Also, preliminary experiments with 18 labels for MIMIC data did not improve the performance of Longformer.

Figure 2 presents the winning F1 score and the differences between the two individual F1 scores for a given label for 93 labels for eICU and 158 labels for MIMIC III data. The best performing (refer to Table 1) embeddings based neural network and transformers for each dataset is represented by different impulses in the Figure 2. Positive F1 scores represent the best F1 score for each label of the two compared systems: Bio-Med-RoBERTa-base and DR-CAML for eICU, and Longformer and CAML for MIMIC III. The negative F1 scores represent the difference between the worst and the best compared F1 scores. Both data

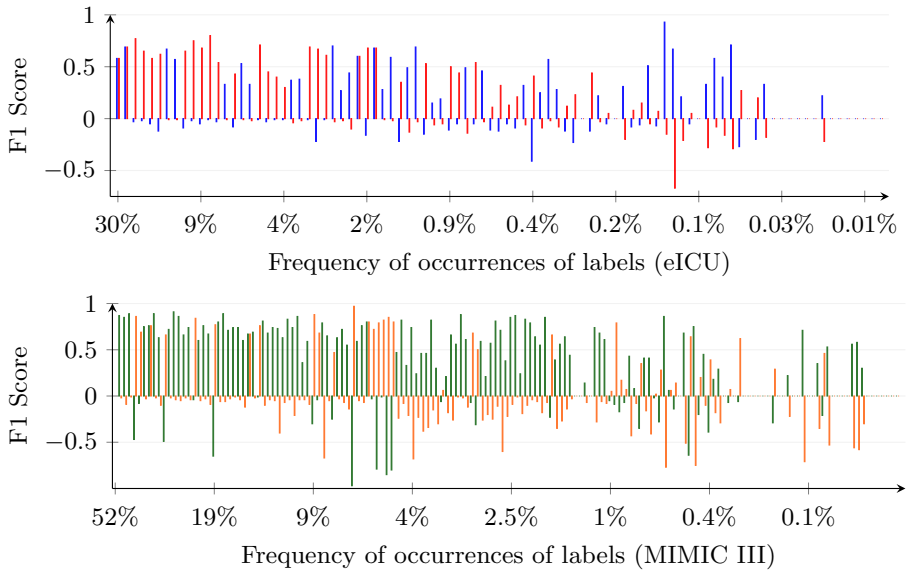

Fig. 2: The winning F1 score for each label, and the difference between the F1 scores from two networks are presented. Best F1 score is represented in the positive y-axis and the difference in the negative y-axis. F1 scores of 93 labels of eICU (top) where ▯▯ is BioMed-RoBERTa-base and ▯▯ is DR-CAML, and 158 labels of MIMIC III (bottom) where ▯▯ is CAML, and ▯▯ is Longformer.

labels are ordered per frequency of occurrence. For eICU, for most labels with frequency $> 0.2\%$ F1 scores obtained using Bio-Med-RoBERTa-base are equal to or better than the DR-CAML ones. In some cases, for label frequencies between 0.7% to 0.2%, F1 scores obtained using DR-CAML are zero, while this is not the case for the transformer model. However, for infrequent labels, DR-CAML has a slight edge over transformer models. For MIMIC III data, for most labels F1 scores obtained using CAML model are better than the Longformer ones. Also, for rare labels the CAML model predicts some labels well, whereas Longformer's F1 scores are mostly zero.

7 Conclusions

This paper has shown that using transformers, especially domain-specific pre-trained models, can be highly beneficial in multi-label medical text classifications. We have presented new SOTA results for predicting medical codes from electronic health records for two very different text datasets, highly pre-processed semi-structured eICU, and free-form MIMIC III, using a fixed sequence length and a number of labels less than or equal to 158. We show that new transformer models, such as Longformer, can be beneficial for long medical documents. Performance is improved compared to standard transformer models, which can only handle sequences of at most 512 tokens.

For longer documents and larger label sets transformers do not show improvements in results when compared to traditional neural networks. Also, imbalanced label distributions are poorly predicted when transformer models are used. Our future works includes looking at ideas such as dual BERT and Siamese BERT to enhance transformers' performance for longer documents. Other research avenues include exploring extreme multi-label classification techniques using transformers such as X-Transformer, and considering medical codes as a hierarchical multi-label problem.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78 (2019)
2. Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K.A., Wixted, M.K.: MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT. In: CLEF (Working Notes) (2019)
3. Amin-Nejad, A., Ive, J., Velupillai, S.: Exploring Transformer Text Generation for Medical Dataset Augmentation. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4699–4708 (2020)
4. Beltagy, I., Peters, M., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
5. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014 (2014)

6. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: ACL (2019)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT (2019)
8. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
9. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779 (2020)
10. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of ACL (2020)
11. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
12. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics (2014)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pretraining distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)
16. Moons, E., Khanna, A., Akkasi, A., Moens, M.F.: A comparison of deep learning methods for icd coding of clinical records. Applied Sciences **10**(15), 5262 (2020)
17. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: 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. ACL: New Orleans, LA, USA (2018)
18. Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data **5**, 180178 (2018)
19. Sänger, M., Weber, L., Kittner, M., Leser, U.: Classifying German Animal Experiment Summaries with Multi-lingual BERT at CLEF eHealth 2019 Task 1. In: CLEF (Working Notes) (2019)
20. Schäfer, H., Friedrich, C.: Multilingual ICD-10 Code Assignment with Transformer Architectures using MIMIC-III Discharge Summaries. In: CLEF 2020 (2020)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30**, 5998–6008 (2017)
22. Yogarajan, V., Gouk, H., Smith, T., Mayo, M., Pfahringer, B.: Comparing High Dimensional Word Embeddings Trained on Medical Text to Bag-of-Words For Predicting Medical Codes. In: Asian Conference on Intelligent Information and Database Systems. pp. 97–108. Springer (2020)