Hierarchical Deep Multi-task learning for Classification of Patient Diagnoses

Salim Malakouti[1][0000−0002−7481−5523] and Milos Hauskrecht[1][0000−0002−7818−0633]

Computer Science Department at University of Pittsburgh
salimm@cs.pitt.edu salimm@cs.pitt.edu

Abstract. Recent years have witnessed an increased interest in the biomedical research community in developing machine learning models and methods that can automatically assign diagnostic codes (ICD) to patient stays based on the information in their Electronic Health Records(EHR). However, despite the recent advances, accurate automatic classification of diagnostic codes continues to face challenges, especially for low-prior diagnostic codes. To alleviate the problem, we propose to leverage information in the diagnostic hierarchy and better utilize the dependencies among diseases in this hierarchy. We develop a new hierarchical deep multi-task learning method that learns classification models for multiple diagnostic codes at the different levels of abstraction in the disease hierarchy while allowing the transfer of information from high-level nodes, more general diagnoses codes to the low-level ones, more specific diagnostic codes. After that, we refine the initial hierarchical model by utilizing the relations and information that can discriminate better between competing diseases. Our empirical results show that our new method and its refinement outperform baseline machine learning architectures that do not leverage the hierarchical structure of target diagnoses tasks or disease-disease relationships.

Keywords: Hierarchical Multi-task Learning · Patient Diagnoses Classification · International Classification Diseases.

1 Introduction

The widespread adoption of electronic health records (EHRs) has introduced the opportunity to process and extract valuable knowledge from massive data warehouses of real-time and diverse clinical data recorded during patient's hospitalizations. One interesting problem is the automatic assignment of diagnostic codes to patients' hospital stays. If the problem is solved successfully, it can help to improve a number of hospital workflows related to both clinical decision-making and administration of healthcare systems. First, diagnostic codes such as the International Classification of Diseases (ICD) are commonly used for hospital reimbursement. The codes are currently assigned to patients by a human annotator (a trained nosologist) after discharge. An effective solution can help to speed up the annotation process and alleviate its cost. Second, an automated diagnostic system could help physicians by providing a concise, automated, and easily accessible summary of patients' conditions and problems not only at the time of discharge but also during the patient's hospital stay. Hence, it can act as a decision support tool that can recommend and bring to the attention of physicians possible patient diagnoses that have not yet been considered. Therefore, recent years have witnessed an increased interest in developing machine learning methods that can automatically assign diagnoses to patient stays based on the information in their electronic health records(EHR) [19, 20, 18]. However, despite recent advancements, multiple challenges making the solutions more practical remain to be solved.

The problem of assigning diagnostic codes to a patient covers is a multi-label or multi-task problem that covers many different diseases. These are organized in various hierarchies or lattice structures, abstracting individual low-level diagnoses into subcategories. This hierarchical structure plays a significant role in the human diagnostic process. Briefly, clinicians are likely able to recognize or reject a high-level diagnostic category much earlier and with a higher certainty than more specific diseases that reside on the lower levels of the hierarchy. Moreover, when the EHR info is incomplete, and information is missing making decisions about some low-level diagnoses may not be feasible. Hence structuring the diagnostic process in a top-down manner based on a hierarchy often helps the clinician to make rapid progress in pursuing feasible diagnoses and arrive at diagnostic conclusions even while additional information is required for a final decision on the most reasonable lower-level assignment. The objective of our work is to bring and leverage the available disease hierarchies into the automatic diagnostic and model learning process. Our conjecture is that machine learning solutions that utilize hierarchies lead to better and more accurate models, and that they can also learn from smaller amounts of available data. The ability to learn models from smaller datasets is important since many low-level diagnoses are rare; that is, they come with a low class prior.

One possible direction for modeling diagnostic code dependencies is multi-task learning. Multi-task learning methods (MTL) have been effective in learning improved machine learning models by facilitating the transfer of knowledge between a set of related target tasks. However, the methods may also fail or are less effective when relations among tasks vary a great deal and when negative transfer in between the tasks may occur [22, 26]. As a result, classic MTL approaches alone may not be sufficient when facing a large number of diagnostic tasks organized in a complex hierarchical structure that can involve task asymmetry and various degrees of task heterogeneity. Therefore, hierarchical multi-task learning methods (HMTL) that can handle hierarchical relations have been proposed to solve an array of problems in natural language processing [23], computer vision [4], speech recognition [9], and even applied to clinical diagnosis problems. However, the limitation of the past hierarchical modeling work for supporting medical diagnoses tasks was somewhat limited and failed to integrate modern deep learning methods to learn more generalizable representations of patients' EHR data sequences. In addition, these methods could not effectively leverage asymmetries and heterogeneity of parent-child relations in existing hierarchies. Our objective in this work is to develop and test new deep learning models and methods across a broad range of diseases organized in the ICD-9 disease hierarchy that leads to improved diagnostic classification models.

We propose a new hierarchical deep learning method that leverages the hierarchical structure of patient diagnoses to facilitate the transfer of information in a top-down fashion, from higher-level diagnostic codes with stronger classification models to lower-level ones. After that, we further refine the initial hierarchical model with a new disease interaction layer. Motivated by the field of differential diagnoses, the interaction layer learns to capture additional patterns from patients' EHR data to better discriminate among competing diagnoses and to fine-tune the predictions of the hierarchical layer. Finally, we compare the performance of our proposed method with baseline algorithms using the MIMIC-III dataset and the ICD-9 diagnosis hierarchy.

2 Related Work

In the following, we briefly review models used for EHR data analysis, solutions for assigning patients instances to diagnoses, and methods for leveraging hierarchies of prediction tasks.

EHR data analysis and automated diagnoses: Most recent work on EHR clinical data analysis and modeling has utilized modern deep learning architectures to learn a low-dimensional representation of patient data based on various NLP model architectures and sequence summaries. These include models based on autoencoders [19], word2vec embeddings and CBOWs summaries [25], recurrent neural networks [2, 20, 12], and various transformer and BERT architectures [13, 21]. These new architectures often lead to improvement of predictive performance over classic featurization methods on a variety of clinical prediction and classification tasks.

One popular application of the above methods was the problem of automatic assignment of diagnoses to patients' EHR sequences [19, 20, 18]. Briefly, Miotto et al. [19] used a denoising autoencoder to generate a low-dimensional representation of the patient state and applied it to multiple clinical classification problems, including patient diagnosis. On the other hand, Rajkomar et al [20], and Lipton et al. [14] used Recurrent Neural Network (RNN) architecture to predict patient discharge diagnoses from a set of clinical variables and sequences of their observations. The diagnostic task was cast as a multi-label classification problem. Finally, Malakouti and Hauskrecht[18] used unsupervised low-dimensional summaries based on SVD to predict patient discharge diagnoses by leveraging a broad range of clinical data (labs, medications, procedures, etc.). The lower-dimensional features were then used to learn independently trained classification models to classify a broad range of patient diagnoses.

While the ML models for supporting the assignment of diagnoses to patients' EHRs have been quite popular, only a limited number of works have tried to leverage and improve the performance of these diagnostic models with the help of disease hierarchies available in ICD-9 and ICD-10 codes. One methodology specifically designed to work with the disease hierarchy was Hierarchical Adaptive MTL (HA-MTL) [17] that relied on an iterative algorithm to learn improved SVM classification models for individual diseases via top-down, and bottom-up sharing of predictive information across the hierarchy [17]. Their results showed that sharing of the predictive information may indeed lead to improved classification models, with top-down sharing accounting for the majority of the improvement. However, we note that the design of these hierarchical methods was somewhat limited and did not use modern EHR sequence embedding methods.

Hierarchical multi-task learning: Multi-task learning (MTL) methods have been proposed to exploit task relationships, their commonalities, and differences to learn improved classification models by allowing transfer of knowledge between the target tasks[27]. In recent years, deep multi-task learning approaches have also shown promising results [3]. Unfortunately, the main shortcoming of early MTL methods is that they relied heavily on the relatedness of target tasks; hence negative transfer could happen when tasks are not sufficiently similar [22]. Various methods have been proposed to prevent negative transfer that leverage underlying task clusters [6, 16], task-task relatedness [1, 8, 15], or facilitate an asymmetric transfer of knowledge [10, 11]. However, neither of these approaches is sufficient to prevent negative transfer when a large number of heterogeneous tasks with various levels of similarities are available. To address these shortcomings, hierarchical multi-task learning methods (HMTL) were introduced [28, 5, 17]. Hierarchical deep MTL methods have also been proposed that leverage the hierarchical structure of a set of carefully selected NLP tasks by allowing inductive transfer of features between task-specific RNN blocks[23]. In computer vision, HD-MTL was proposed, which first learned a visual tree for a large set of atomic object classes and then leveraged the inter-class relatedness in the visual tree to jointly learn more representative deep CNNs and a more discriminative tree classifier for the target tasks[4]. However, to the extent of our knowledge, this work will be the first attempt to propose a deep HMTL method that leverages the diagnoses hierarchy to promote a top-down transfer of features from parents to children while modeling the interactions between closely related tasks at the same level of the hierarchy (siblings). Additionally, our proposed method is the first attempt to also incorporate disease-disease interactions of sibling diagnoses to learn improved diagnostic classifiers.

3 Methodology

Let D be the number of target diagnostic tasks of varying difficulty organized in a hierarchical structure H. Our goal is to learn classification models for each of these tasks by taking advantage of task relations reflected in H.

The patients' EHRs are formed by complex sequences of observations, physiological events, treatments, and procedures. To facilitate the learning of classification models, the EHR sequences are often replaced with a compact vector-based representation that attempts to summarize the information in EHRs relevant to the specific prediction tasks. This transformed representation is also referred to as embedding. In the following, we first describe the basic architecture for transforming the data in EHRs to a lower-dimensional embedding. After that, we propose a refinement of this architecture by incorporating a hierarchical multi-task learning layer that facilitates sharing of embeddings among related tasks. Finally, we add a new model layer that incorporates disease-disease interactions to learn additional task-specific features that aim to further refine the different diagnostic models.

3.1 EHR data pre-processing and initial EHR transformation

We have adopted a three-step process to generate initial patient representations from a wide range of patient clinical data (Figure 1). In Step 1, binary events representing lab results, medications, vital signs, and procedures are generated from the patient's raw medical records. In Step 2, the events are divided into T segments of equal length (24h window size). The events in each time segment are then represented by normalized Bag-of-Words (BoW) vectors. Finally, the normalized BoW vectors for each segment are then transformed using a supervised feed-forward layer (Figure 2: Embedding Layer). The weights of the feed-forward layer are learned from available data, and its output defines the initial EHR embedding vector v_t where $t \in \{1, 2, .., T\}$. Please note there is one embedding vector v_t per segment of time per patient.

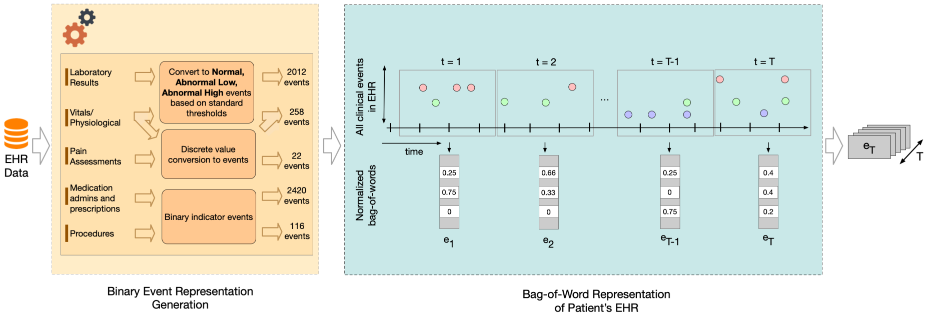


Fig. 1: Preprocessing steps to learn lower-dimensional representation of EHR

3.2 Hierarchical Multitask Learning Layer

Multi-task learning aims to train target tasks simultaneously and, hence, learn improved classification models by facilitating the transfer of knowledge between related tasks. In deep multi-task learning methods, this similarity is often achieved through either a set of common latent feature layers shared by all or groups of related tasks or through imposed similarities between a set of task-specific constrained feature layers. However, traditional methods may fail to efficiently leverage task relationships when facing a large number of heterogeneous tasks with various levels of similarities. There, hierarchical MTL methods aim to leverage underlying task hierarchies to efficiently direct sharing of information between target tasks.

Our proposed layer learns a separate set of task-specific neural network blocks for each target task in any arbitrary hierarchy while facilitating the inductive transfer of features in a top-down fashion by sharing hidden states of parent tasks with its children (see Figure 2). Additionally, following Sanh et al. [23] we use shortcuts (blue arrows) so that each target task can have access to the original EHR feature embeddings. This dual input mechanism enables each target task to either learn new features from the shared EHR embeddings, adopt features from more general categorical parent tasks p (black arrows), or combine these two sets of features in order to learn improved classification models. This is analogous to clinicians distinguishing specific diagnoses types by examining additional information that helps identify them from the other members of a group of diseases with similar symptoms. Task-specific blocks in this work are modeled using a bi-directional LSTM encoder architecture. The encoders take as input the concatenated vector of original EHR embeddings (v_t vectors) and the hidden states of their parent task p at each timestamp t (h_t^p). Next, a max-pooling layer($max([h_1^m, h_2^m, ..., h_T^m])$) for each target task was adopted to combine task-specific LSTM hidden states at all timestamps. Finally, a feed-forward layer with a sigmoid activation function was adopted to learn the final classification scores for each target task.

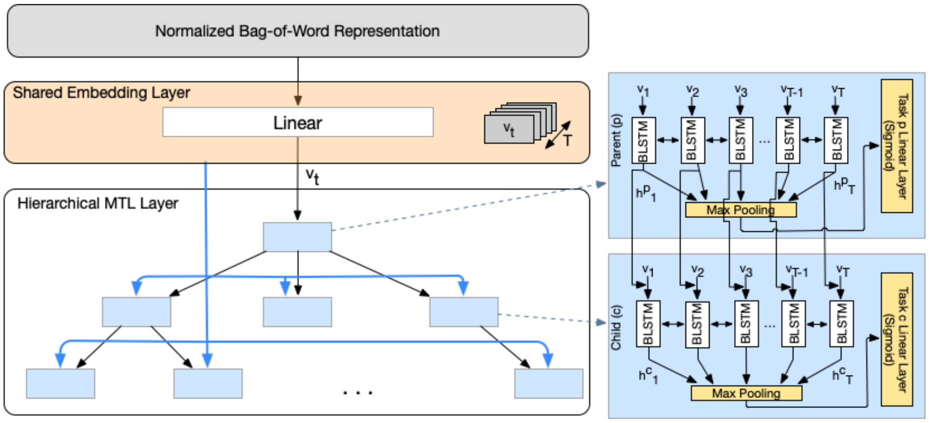
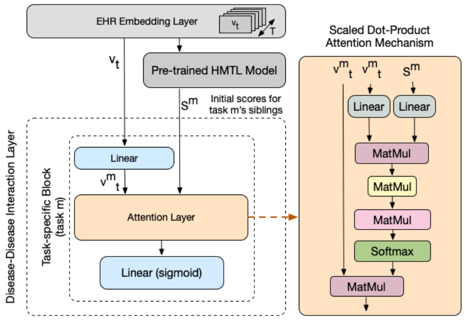


Fig. 2: The proposed HLSTM network architecture.

3.3 Disease-Disease Interaction Layer

Differential diagnoses in medicine refer to distinguishing a particular patient's disease from a set of competing diagnoses with similar features through systematic methods of acquiring and examining additional data. Similarly, a comprehensive machine learning solution should capture such disease-disease interactions to classify patients' diagnoses accurately. Therefore, we propose a fine-tuning step that is trained separately as a second step and learns to capture additional patterns from patients' EHR data to improve the initial predictions by the hierarchical layer. The interaction layer, defines the final prediction probability for the target task m as $\hat{f^m} = sigmoid(f^m + \Delta f^m)$ where f^m is the initial score based on the hierarchical model and Δf^m determines the change to the scores based on the disease-disease interactions with its siblings. Motivated by the field of differential diagnoses, a task-specific feature attention-based learning block is adopted to learn additional features (Figure 3). First, a single linear layer is used to learn a low-dimensional task-specific feature vector v_t^m from the original EHR embeddings v_t for each target task m. This is followed by a scaled dot-product attention layer similar to the multi-head attention mechanism proposed in "Attention is All You Need"[24] that uses v_t^m vectors and the initial classification scores S^m from task m's siblings to learn a set of importance weights α_t^m for each timestamp t. Finally, a final feature vector is obtained as $v^m = \sum_t^T \alpha_t^m v_t^m$. Please note that this task-specific architecture uses the initial predictions of siblings and the original EHR embeddings to capture new information from the most important window segments during a patient's hospitalization to fine-tune the initial predictions.



$$\begin{aligned} \hat{f^m} &= sigmoid(f^m_{hmtl} + \Delta f^m) \\ \Delta f^m &= W_s V^m + b_s \\ V^m &= attn(W_q S^m, W_k V^m_T, V^m_T) \end{aligned} \tag{1}$$

while:

$$attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{h}})V \tag{2}$$

Fig. 3: Task specific interaction layer

4 Experiments

Data Description: The experiments in this section are conducted using the MIMIC-III dataset [7], an open-access EHR dataset collected over a 12-year time span. Only patients included in the MetaVision subset, including 22,046 visits, were included since the coding terminology used for patients' clinical data has a higher coherency. Finally, ICD-9 discharge diagnostic codes were used to create diagnostic labels. Since medical diseases are only recorded using the lower level leaf diagnostic codes, binary labels for the diagnostic categories were created by applying a logical OR operation between all its children. Finally, we consider only diagnoses and diagnostic categories that satisfy a minimum cut-off threshold($N_{min} = 100$) on the number of patients with that diagnosis (D = 1228) to ensure sufficient positive sample sizes.

Implementation Details: The proposed HLSTM architecture was implemented with a linear embedding layer of dimension 256 and the task specific bi-LSTM used a hidden state of size 32. For evaluation, we adopted the weighted area under the receiver operating curve (AUROC) and the area under the precision-recall curve (AUPRC), which is suggested to be more suitable when using the average of the metrics across multiple imbalanced target tasks with varying skewedness[20]. Finally, a random split of (70%/30%) the data was generated to create train and test sets.

Overall performance: In this section, we compare the overall performance of our proposed method with baselines including: (1) multi-label LSTM architecture which included a bidirectional LSTM layer, followed by a max-pooling layer and a linear layer with sigmoid activation function to classify all target tasks (multi-label lstm), (2) and MTL lstm implementation that adopted a shared feature embedding layer as the common feature layer while using task-specific lstm blocks that share information through the shared embedding layer. All baselines utilize the same EHR feature learning method as HMTL and use mean binary cross-entropy loss for training.

Our empirical results show that our HMTL method results in strong improvements across all tasks, while the interaction layer also introduces slight improvements over the HMTL layer. These improvements are consistent among both categorical and low-level leaves (low-prior and imbalanced), showing that the proposed method was able to transfer information top-down in an effective manner.

	All Nodes		**Category Nodes**		**Leaf Nodes**	
Method Name	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Multi-label lstm	0.76	0.74	0.752	0.735	0.766	0.742
MTL lstm	0.69	0.674	0.724	0.70	0.675	0.653
HMTL	0.805	0.799	0.801	0.796	0.808	0.80
HMTL + Interaction layer	0.817	0.803	0.806	0.801	0.815	0.806

Table 1: Comparison of overall performance of proposed method with baselines (average AUROC and AUPRC)

Task level analysis: While the overall results show strong improvements across all diagnoses and diagnostic categories ($M = 1228$), it's still valuable to evaluate the performance of the model across individual tasks. Figure 4 shows improvements in the individual target diagnostic tasks with respect to both weighted AUROC and weighted AUPRC metrics. In general, our proposed method resulted in considerable improvement ($\Delta > 0.05$) of nearly 50% of target tasks while preventing negative transfer with more 91% of classifiers performing at least as good as the baseline models($\Delta \geq 0$). In fact, only a handful of very rare diagnoses(2% of $0.004 \geq prior < 0.01$ group) demonstrated considerably lower performance that the baseline models ($\Delta < -0.05$). While a perfect MTL method is expected to only result in positive improvements, this has proven difficult in practice, especially when facing a large number of target tasks[26]. We conjecture that the negative improvements are mainly due to the imperfect hierarchy designs caused by residual categories that include diagnoses not closely aligned with other diseases. This motivates research and development of future HMTL methods that simultaneously learn to improve the existing hierarchies for machine learning tasks.

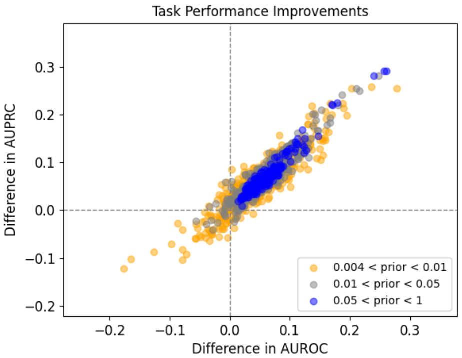


Fig. 4: Performance improvements of individual tasks compared to the baseline multi-label LSTM models

5 Conclusion

We propose a new hierarchical deep learning method that leverages the hierarchical structure of patient diagnoses to allow the transfer of information in a top-down fashion, from higher-level diagnostic codes with stronger classification models to lower-level ones. After that, we refine the initial hierarchical model with a new disease interaction layer, utilizing the task relationships and new patient information to learn classifiers that can better discriminate between competing diagnoses. Our results show that our proposed method strongly outperforms baselines across all target tasks, resulting in positive transfer in nearly 50% and preventing negative transfer in 92% of the target diagnoses.

References

1. Ben-David, S., Schuller, R.: Exploiting task relatedness for multiple task learning. In: Learning Theory and Kernel Machines, pp. 567–580. Springer (2003)
2. Choi, E.e.a.: Mime: Multilevel medical embedding of electronic health records for predictive healthcare. arXiv preprint arXiv:1810.09593 (2018)
3. Crawshaw, M.: Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796 (2020)
4. Fan, J., Zhao, T., Kuang, Z., Zheng, Y., Zhang, J., Yu, J., Peng, J.: Hd-mtl: Hierarchical deep multi-task learning for large-scale visual recognition. IEEE transactions on image processing **26**(4), 1923–1938 (2017)
5. Han, L., Zhang, Y.: Learning tree structure in multi-task learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 397–406. ACM (2015)
6. Jacob, L., Vert, J.p., Bach, F.R.: Clustered multi-task learning: A convex formulation. In: Advances in neural information processing systems. pp. 745–752 (2009)
7. 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)

8. Kang, Z., Grauman, K., Sha, F.: Learning with whom to share in multi-task feature learning. In: ICML. vol. 2, p. 4 (2011)
9. Krishna, K., Toshniwal, S., Livescu, K.: Hierarchical multitask learning for ctc-based speech recognition. arXiv preprint arXiv:1807.06234 (2018)
10. Lee, G., Yang, E., Hwang, S.: Asymmetric multi-task learning based on task relatedness and loss. In: International Conference on Machine Learning (2016)
11. Lee, H.B., Yang, E., Hwang, S.J.: Deep asymmetric multi-task feature learning. In: International Conference on Machine Learning. pp. 2956–2964. PMLR (2018)
12. Lee, J.M., Hauskrecht, M.: Modeling multivariate clinical event time-series with recurrent temporal mechanisms. Artificial Intelligence in Medicine **112** (2021)
13. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: transformer for electronic health records. Scientific reports **10**(1), 1–12 (2020)
14. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with lstm recurrent neural networks. arXiv preprint (2015)
15. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
16. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: IEEE conference on computer vision and pattern recognition (2017)
17. Malakouti, S., Hauskrecht, M.: Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE (2019)
18. Malakouti, S., Hauskrecht, M.: Predicting patients diagnoses and diagnostic categories from clinical-events in ehr data. In: AIME (2-22)
19. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports **6**, 26094 (2016)
20. Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., et al.: Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine **1**(1), 1–10 (2018)
21. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 1–13 (2021)
22. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer. In: NIPS 2005 workshop on transfer learning. vol. 898, pp. 1–4 (2005)
23. Sanh, V., Wolf, T., Ruder, S.: A hierarchical multi-task approach for learning embeddings from semantic tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6949–6956 (2019)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
25. Zhang, J., Kowsari, K., Harrison, J.H., Lobo, J.M., Barnes, L.E.: Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access **6**, 65333–65346 (2018)
26. Zhang, W., Deng, L., Zhang, L., Wu, D.: Overcoming negative transfer: A survey. arXiv preprint arXiv:2009.00909 (2020)
27. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)
28. Zweig, A., Weinshall, D.: Hierarchical regularization cascade for joint learning. In: International Conference on Machine Learning. pp. 37–45 (2013)