A Unified Review of Deep Learning for Automated Medical Coding

SHAOXIONG JI, Aalto University, Aalto, Finland
XIAOBO LI, Dalian Maritime University, Dalian, China
WEI SUN, KU Leuven, Leuven, Belgium
HANG DONG, University of Exeter, Exeter, United Kingdom of Great Britain and Northern Ireland
ARA TAALAS, Terveystalo Healthcare Services, Helsinki, Finland
YIJIA ZHANG, Dalian Maritime University, Dalian, China
HONGHAN WU, University of Glasgow, Glasgow, United Kingdom of Great Britain and Northern Ireland
ESA PITKÄNEN, University of Helsinki, Helsinki, Finland
PEKKA MARTTINEN, CS, Aalto University, Aalto, Finland

Automated medical coding, an essential task for healthcare operation and delivery, makes unstructured data manageable by predicting medical codes from clinical documents. Recent advances in deep learning and natural language processing have been widely applied to this task. However, deep learning-based medical coding lacks a unified view of the design of neural network architectures. This review proposes a unified framework to provide a general understanding of the building blocks of medical coding models and summarizes recent advanced models under the proposed framework. Our unified framework decomposes medical coding into four main components, i.e., encoder modules for text feature extraction, mechanisms for building deep encoder architectures, decoder modules for transforming hidden representations into medical codes, and the usage of auxiliary information. Finally, we introduce the benchmarks and real-world usage and discuss key research challenges and future directions.

CCS Concepts: • **Applied computing** → **Health care information systems**; Document management and text processing; • **Computing methodologies** → *Natural language processing*.

Additional Key Words and Phrases: Medical Coding, Deep Learning, Unified Framework

1 INTRODUCTION

In the field of Natural Language Processing (NLP), deep learning that builds deep neural networks for representation learning has attracted significant attention from the research community and achieved superior performance in various applications such as automatic extraction of useful information and answer generation from human queries [139, 167]. NLP techniques allow the machine to process human languages automatically and have been widely studied in analyzing health-related texts, measuring healthcare quality, and promoting the delivery of healthcare services. For example, Sentic PROMs (patient-reported outcome measures) enable patients' physio-emotional sensitivity tracking and measuring healthcare quality through sentic computing on

Authors' Contact Information: Shaoxiong Ji, Aalto University, Aalto, Finland; e-mail: shaoxiong.ji@helsinki.fi; Xiaobo Li, Dalian Maritime University, Dalian, Liaoning, China; e-mail: xiaobo.li@dlmu.edu.cn; Wei Sun, KU Leuven, Leuven, Belgium; e-mail: wei.sun@aalto.fi; Hang Dong, University of Exeter, Exeter, United Kingdom of Great Britain and Northern Ireland; e-mail: h.dong2@exeter.ac.uk; Ara Taalas, Terveystalo Healthcare Services, Helsinki, Uusimaa, Finland; e-mail: ara.taalas@terveystalo.com; Yijia Zhang, Dalian Maritime University, Dalian, Liaoning, China; e-mail: zhangyijia@dlmu.edu.cn; Honghan Wu, University of Glasgow, Glasgow, United Kingdom of Great Britain and Northern Ireland; e-mail: honghan.wu@ucl.ac.uk; Esa Pitkänen, University of Helsinki, Helsinki, Uusimaa, Finland; e-mail: esa.pitkanen@helsinki.fi; Pekka Marttinen, CS, Aalto University, Aalto, Finland; e-mail: pekka.marttinen@aalto.fi.

ACM 1557-7341/2024/5-ART
https://doi.org/10.1145/3664615

free-text patient notes [21]. Contextualized text representations and classification models also facilitate outbreak management during epidemics [90]. There are other patient-centered applications, such as proactive mental healthcare [79] and patient opinion mining [22], to name a few. This review focuses on deep neural network-based NLP techniques for automated medical coding with medical ontologies, also known as medical code assignment, medical code prediction, medical coding, or clinical coding. Medical code assignment uses all types of clinical notes to predict medical codes in a supervised manner with human-annotated codes [144], formulated as a multi-class multi-label text classification problem in the medical domain. Most deep learning-based medical coding models are trained in a centralized manner, while some recent publications investigate the emerging application of federated learning [34].

Healthcare workers write clinical notes about a patient's health status to document their insights and observations for further diagnosis decision support. Clinical notes as free-text descriptions are an essential component of Electronic Health Records (EHRs), which contain patient medical history, symptom description, lab test result summary, reasons for diagnoses, and daily activities [105]. Diagnosis codes in typical medical classification systems identify a patient's diseases, disorders, symptoms, and specific reasons for the hospital visit. In contrast, procedural codes or intervention codes identify surgical, medical, or diagnostic interventions. Diagnosis codes, which a trained health professional assigns, act as the standard translation of written patient descriptions. Diagnostic coding is an integral part of the clinical coding process in health information management with procedural codes. Medical coding, particularly for billing purposes, often operates separately from the clinical care process and may not significantly impact immediate patient care decisions. While medical coding primarily serves billing and administrative purposes, accurate coding does have broader implications for healthcare quality assessment, research, and resource allocation [18, 20, 132]. It can play a role in retrospective analysis, identifying trends in patient populations, and assessing the effectiveness of certain treatments or interventions.

Clinical notes are usually annotated with standardized statistical codes to facilitate information management. Different diagnosis classification systems utilize various medical coding systems. The International Classification of Diseases (ICD) system, maintained by the World Health Organization (WHO), is one of the most widely-used coding systems adopted in countries across the globe[1]. The ICD system transforms diseases, symptoms, signs, and treatment procedures into standard medical codes. It has been widely used for clinical data analysis, automated medical decision support [37], billing, and medical insurance reimbursement [141]. Specific versions of ICD include ICD-9, ICD-9-CM, ICD-10, and ICD-11. Most ICD-9 codes consist of three digits to the left of a decimal point and one or two digits to the right. Some ICD-9 codes have "V" or "E" in front of the digits, representing preventive health services and environmental causes of health problems. Figure 1 shows a fragment of the patient's clinical note with ICD-9-CM codes assigned. The ICD-9-CM created by the US National Center for Health Statistics (NCHS) adapts ICD-9 codes used in the United States. The first three characters of ICD-10 codes define the category, and the next three digits describe the etiology, anatomic site, severity, and other vital information. The latest ICD version is ICD-11, which will become effective in 2022, while older versions such as ICD-9, ICD-9-CM, and ICD-10 are also concurrently used. Other widely used medical condition classification systems include the Clinical Classifications Software (CCS) and Hierarchical Condition Category (HCC) coding. It is worth noting that CCS and HCC are derived directly from ICD codes and serve as specialized, broader categorizations within the ICD framework, tailored for specific purposes.

Properly coded medical information is vital for clinical decision-making, public health surveillance, research, and reimbursement. Automated care pathways are often triggered by patients receiving a specific diagnosis code. On the national scale, care guidelines are often structured around diagnosis codes, providing interventions for clearly defined conditions [54, 186]. On the healthcare provider side, quantitative measurement of healthcare effectiveness and care development is, by necessity, based on code-based logic. Questions, such as how many

[1]https://www.who.int/standards/classifications/classification-of-diseases

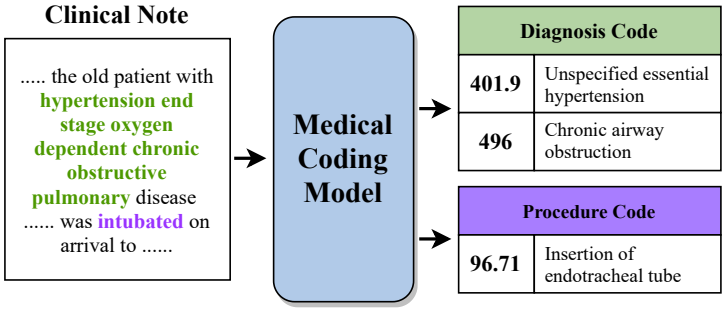


Fig. 1. A medical coding model maps an example clinical note to the corresponding ICD procedure and diagnosis codes.

patients diagnosed with a given illness received appropriate care, can only be measured quantitatively based on medical coding [47]. Automated diagnosis coding can also be deployed to detect missed diagnoses and adverse effects [126]. Effective treatment often relies on the early detection of symptoms, and pre-emptive healthcare can only be built on technology sensitive to slight deviations in the patient's health state.

Automatic medical code assignment uses feature engineering techniques and machine learning-based classifiers to predict medical codes from clinical notes [41]. Medical coding requires efficient matching between textual mentions and specific diagnosed codes. It exploits the dependencies between input and output variables by learning structured output representation. Traditional medical coding systems deploy rule-based methods [59], select manual features [125], and apply machine learning-based classification models such as Support Vector Machines (SVM) [17] and Bayesian ridge regression [112]. The hierarchical structure of the code system, depicted as a tree structure with multiple levels, is a basic pattern for improving the automated coding method. For example, Perotte et al. [144] adopted the ICD hierarchy and developed flat and hierarchical SVM for diagnosis code classification. The breakthrough of natural language processing with deep neural networks has led to neural classifiers with word embedding and deep learning [182]. Neural methods for medical text encoding intensively use recurrent neural networks (RNNs), convolutional neural networks (CNNs), and neural attention mechanisms, where parameter selection is a vital issue [113].

However, three main challenges remain in processing medical text and automated coding.

(1) Noisy and Lengthy Clinical Notes. Clinical notes contain many professional medical vocabularies and noisy information such as non-standard synonyms and misspellings. They are usually lengthy documents containing many types of clinical information, such as health profiles, lab tests, radiology reports, operative reports, and medications. Thus, they typically have hundreds or even thousands of words. Some patients with long hospital stays may have much longer written notes. Additionally, writing styles can vary from one healthcare professional to another, with domain-specific lingo giving a given word different meanings depending on context. The medical practice also evolves with time, with coding systems and notation changing from one year to another.

(2) High-dimensional Medical Codes. Medical notes are associated with multiple diagnoses, usually treated as a multi-label extreme classification problem containing a large label set. The high-dimensional label space has thousands of codes. For example, ICD-9 and ICD-10 coding systems have more than 14,000 and 68,000 codes, respectively. The space of target classes is exponential to the number of output classes making it extremely challenging when facing high-dimensional medical ontologies.

(3) Imbalanced Classes. A patient typically is diagnosed with only a couple of codes over the whole coding space, while patients with complicated diseases are associated with dozens of codes. Moreover, because of the

existence of common and rare diseases, the distribution of medical codes in an EHR system is imbalanced, also known as the long-tail phenomenon. For example, the distribution is highly skewed in the MIMIC-III dataset [85], as shown in Fig. 2[2]. The limitation of data acquisition also exacerbates the imbalance. The data from intensive care units such as the MIMIC-III contains severe cases with other complications. Patient records of the visit to general practitioners only have some general codes.

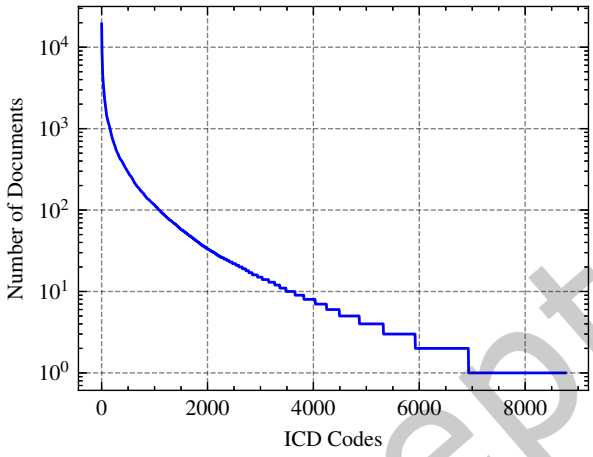


Fig. 2. The distribution of ICD codes in the MIMIC-III dataset curated by Mullenbach et al. [131].

As an interdisciplinary study, the successful development of automated medical coding requires the collaboration between computer scientists and clinical coders [172]. Automated medical coding still has a long way to go with such challenging matters. This review illustrates the development trend in recent deep learning-based medical coding methods and proposes a unified encoder-decoder framework to shed light on future research. We investigate how the existing methods can be categorized into the encoder-decoder framework widely adapted by many AI applications and summarize and review deep learning-based natural language processing techniques for automated medical coding to provide theoretical and pragmatic insights into the varieties and nuances of neural network architectures in this field. We unify recent neural network-based methods into an encoder-decoder framework and introduce them under this unified framework. This review is organized as follows. Sec. 2 introduces related reviews in this field and highlights our contributions. We formulate the unified encoder-decoder framework and corresponding building components in Sec. 3. Sec. 4 introduces widely-used benchmarks and real-world applications. We discuss current limitations and point out future research directions in Sec. 5. Finally, we conclude our studies in Sec. 6.

2 RELATED REVIEWS AND CONTRIBUTIONS

There have been several systematic and narrative reviews on automated medical coding, as summarized in Table 1. One of the first reviews reported the published accuracy of discharge coding in literature [24]. Burns et al. [19] conducted an updated review on the accuracy of routinely collected data following Campbell et al. [24]. Stanfill et al. [160] introduced some conventional classification methods and evaluated different types of automated coding systems. Campbell et al. [23] conducted an application-oriented review of computer-assisted clinical

[2]The number documents on the y-axis, rather than the number of visits, follows the convention of Mullenbach et al. [131] and is mostly used in NLP community.

coding. A recent systematic review [88, 89] followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and searched publications with machine learning (ML) and natural language processing techniques. The authors reviewed publications from 2010 to 2020 in a narrative way. Khope and Elias [91] conducted a similar systematic review but focused on the studies that used the MIMIC-III dataset. Fewer reviews covered technical matters in medical coding. Teng et al. [164] conducted a technical review that discusses recent advances in machine learning and natural language processing on medical coding. Published in early 2022, it is a concurrent work focusing on feature engineering-based classifiers and deep learning methods. However, it does not provide a unified view that can generalize to all the nuanced varieties of deep learning architectures or cover the most recent learning algorithms for automated medical coding. Those limitations of exisiting reviews motivate us to propose a unified view of automatic medical coding models.

Publications	Period	Category	Scope or Focus
Campbell et al. [24]	1975 - 1998	Systematic review	Coding accuracy
Stanfill et al. [160]	1996 - 2009	Systematic review	Automated coding tools
Burns et al. [19]	1990 - 2010	Systematic review	Coding accuracy
Campbell et al. [23]	2006 - 2017	Narrative review	Computer-assisted clinical coding
Kaur et al. [88]	2010 - 2020	Systematic review	ML and NLP techniques
Teng et al. [164]	1990s - 2021	Technical review	ML and NLP techniques
Khope and Elias [91]	2017-2023	Systematic review	Studies on MIMIC-III dataset
Kaur et al. [89]	2010 - 2021	Systematic review	ML and NLP techniques
Ours	2010s - 2023	Technical review	Unified deep learning framework

Table 1. A summary of related review articles on medical coding

Previous reviews introduce conventional classification systems or neural network-based methods that devise various network architectures to improve predictive performance. However, there is no unified study on medical coding models nor an insightful analysis of the overall model architecture's submodules to solve the challenges mentioned earlier. Besides, recent deep learning advances beyond standard supervised learning are less discussed in existing surveys. The emerging deep learning paradigms include multitask learning, few-shot and zero-shot learning, contrastive learning, adversarial generative learning, and reinforcement learning.

This paper focuses on deep learning-based NLP techniques and proposes an encoder-decoder framework (Fig. 3) to unify existing advanced medical coding models. It discusses the effect of different building blocks to resolve the challenges of medical coding. The categorization of building blocks is summarized in Table 2. We provide a complete guideline for researchers or practitioners to develop efficient neural networks for automated medical coding and analyze the critical problems for tackling the existing challenges. Besides, we discuss the evaluation of medical coding and its real-world practice. Finally, we summarize the recent research trends and limitations and point out several vital directions for future research. Medical coding tasks evolve rapidly, with many deep learning-based publications emerging. In this review, we curate a collelction of publications that generally come from academic databases such as PubMed, IEEE Xplore, and ACL Anthology. We conduct this timely review to fill the gap by presenting a unified review and introducing recent advances in deep neural architectures for automated medical coding and emerging learning paradigms beyond supervised learning.

The development of a unified encoder-decoder framework for advanced medical coding models, as discussed in this paper, serves several crucial purposes in the field of NLP applied to healthcare. This framework serves to unify existing models, tackle specific challenges in medical coding, and offer practical guidelines for researchers and practitioners. It emphasizes the importance of real-world evaluation and staying current with the rapidly evolving landscape of medical coding tasks. By summarizing recent trends, identifying limitations, and suggesting future

research directions, the paper provides a comprehensive resource for advancing the development of efficient neural networks for automated medical coding and supports the continued growth of this critical field.

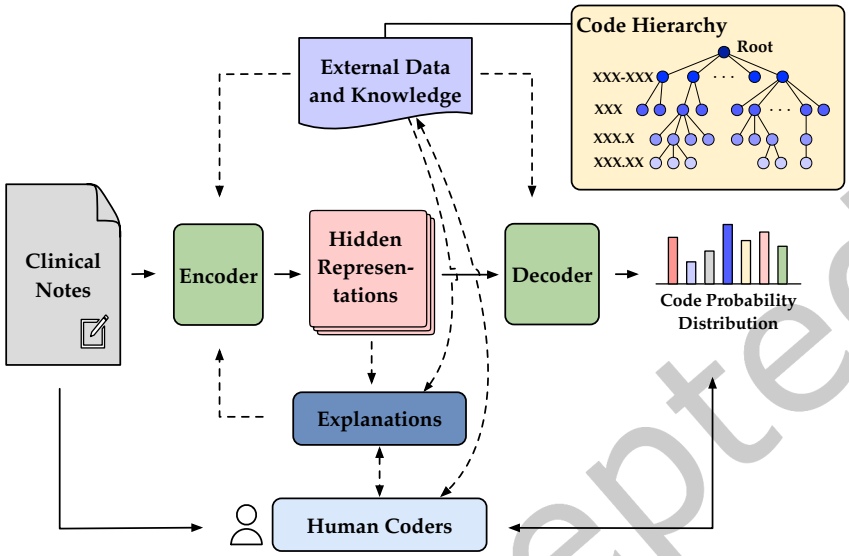


Fig. 3. An illustration of the unified encoder-decoder framework for automated medical coding

Categories	Functions	Representative Methods
Encoders	Extract text features, explainable feature learning	CNN, RNN, graph neural networks, attention, Transformers, capsule networks
Deep Connections	Build deep architecture	Stacking, residual networks, embedding injection
Decoders	Improve code prediction	Linear layer, attention, hierarchical decoders, multitask decoders, few-shot/zero-shot decoders, autoregressive generative decoders
Auxiliary Data	Enhance feature learning, human-in-the-loop learning	Code descriptions, code hierarchy, Wikipedia articles, chart data, entities and concepts, human-in-the-loop learning

Table 2. Categorization of building blocks under the unified framework

3 A UNIFIED ENCODER-DECODER FRAMEWORK

The recent development of automated medical coding devises novel neural networks for medical code prediction. For example, recurrent neural network (RNN) based methods such as the long short-term memory (LSTM) network with attention mechanism [157] and GRU network with hierarchical attention [11] have been widely applied to medical code prediction from discharge summaries. Deep learning-based CNN models have been compared with a conventional classifier for diagnosis coding from radiology reports [87]. In addition to conventional supervised learning, many novel learning paradigms have also been studied, for example, multitask learning [210] and few-shot learning [184]. This review focuses on deep learning-based NLP techniques applied to automated

medical coding, unifies recent advances to introduce their advantages, and summarizes their building blocks' theoretical and pragmatic motivations.

We propose a unified encoder-decoder framework (Fig. 3) for automated medical coding. In particular, encoders refer to modules responsible for extracting relevant features from clinical text data, while decoders transform these features into medical codes. The encoder modules take clinical notes as inputs and learn hidden representations, as described in Sec. 3.1. One important aspect of hidden representation learned by neural encoders is to produce explanations and enable trustworthy coding systems grounded by human evaluation. Machine learning (ML) algorithms excel in prediction and performance but often lack transparency, emphasizing the need for explainable systems as discussed by Adadi [1]. In the ML context, "interpretability" signifies a model's inherent understandability, while "explainability" refers to methods to make a model interpretable. Explanations are the specific insights provided by a model to aid users in understanding predictions. A current debate revolves around whether attention mechanisms contribute to model explanation, explored by Bibal et al. [13]. We also introduce and summarize mechanisms for deepening the architectures in Sec. 3.2. The decoder modules decode the hidden representations to predict the code probability (Sec. 3.3). While the choice of encoder and decoder can be interrelated, separating these modules within the framework enhances modularity, flexibility, comparative analysis, interchangeability, and conceptual clarity. These benefits support a more comprehensive exploration of deep learning approaches for medical coding while accommodating the complexity of encoder-decoder interactions in various coding scenarios.

In addition to decoders in the standard supervised setting, we review recent advances such as multitask decoders, few-shot/zero-shot decoders, and autoregressive generative encoders. During encoding and decoding, auxiliary information such as code hierarchy and textual descriptions can also be applied for enhancing representation learning and improving decoding, which is discussed in Sec. 3.4. Besides, augmented learning with external information, especially ontological knowledge, promotes explainable medical coding. Human-in-the-loop learning integrates human coders into the automated medical coding system to further enhance the encoder-decoder framework. They collaborate with automated algorithms to validate and improve coding accuracy. Human coders are involved in various tasks, including validating the suggestions made by automated systems, resolving ambiguous cases, ensuring compliance with coding guidelines, and refining the overall coding process. Their expertise helps refine machine learning models, contributing to ongoing system enhancement and ensuring high-quality, accurate medical coding. For example, active learning reduces annotation costs, and human-grounded evaluation enables reliable performance evaluation. We summarize recent representative models in Table 3 under the proposed unified framework and review them in the following subsections.

3.1 Encoder Modules

Deep learning-based models use word embedding techniques and develop complex neural network architectures to learn rich text features for automatic medical code assignment. After some text preprocessing techniques, a clinical note with n words is denoted as $\{x_0, \ldots, x_n\}$. Its word embedding matrix, for example, built by word2vec [128] or GloVe [143], is denoted as $\mathbf{X} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]^T \in \mathbb{R}^{n \times d_e}$, where d_e is the dimension of word vectors. The encoder modules of various neural architectures further process embeddings to learn rich hidden representations. In the context of deep learning for medical coding, CNN-based approaches focus on extracting local features and patterns from clinical documents, while RNN-based methods are employed to capture sequential dependencies and contextual information within the text. Recent publications also leverage advanced language models like BERT, which provide contextualized word embeddings, enhancing the understanding of medical narratives by considering the broader context of each word in the document. These diverse techniques contribute to more accurate and comprehensive automated medical coding systems. This section introduces various neural encoder modules that have been developed in recent years.

Models	Encoders	Deep Connections	Decoders	Auxiliary Data
Attentive LSTM [157]	Attentive LSTM	Stacking	Linear Layer	NA.
HA-GRU [11]	Hierarchical GRU	Stacking	Attention	NA.
LAAT [173]	BiGRU	Stacking	Attention	NA.
MT-RAM [161]	BiGRU	RAM	LAN+Multitask	NA.
BiCapsNetLE [10]	BiLSTM+CapsNet	Stacking	Attention	ICD Description
CAML [131]	CNN	Stacking	LAN	NA.
DR-CAML [131]	CNN	Stacking	LAN	ICD Description
MVC-LDA [153]	Multi-view CNN	Stacking	Attention	ICD Description
MultiResCNN [104]	CNN	Residual Network	LAN	NA.
DCAN [77]	Dilated CNN	Residual Network	LAN	NA.
HyperCore [26]	CNN+Hyperbolic	Stacking	LAN+GCN	ICD Hierarchy
GatedCNN-NCI [82]	Gated CNN	Embedding Injection	NCI	ICD Description
Fusion [123]	Compressed CNN	Residual Network	Attention	NA.
C-MemNN [147]	Memory Networks	Stacking	Linear Layer	NA.
KSI [9]	CNN/RNN	Stacking	Linear or LAN	Wikipedia Articles
MCDA [183]	CNN/RNN	Stacking	Concept-drive Attention	Wikipedia Articles
MSATT-KG [193]	CNN+Attention	Stacking	Attention+KG	ICD Hierarchy
CAIC [165]	CNN/RNN	Stacking	Attention	ICD Description
GMAN [203]	GCN	Stacking	Mutual Attention	Patient Info.
JLAN [106]	BiLSTM	Residual Network	Self-attention+LAN	ICD Description
DACNM [27]	Dilated CNN	Stacking	N-gram+Linear	ICD Description
BERT-XML [211]	BERT	Stacking	LAN	ICD Description
ISD [214]	CNN	Stacking	Attention	NA.
CMGE [188]	Graph encoder	Stacking	Multitask Decoder	NA.
RAC [93]	CNN	Stacking	Attention	Data Augmentation
ICDBigBird [127]	BigBird	Stacking	Label Attention	NA.
HieNet [181]	CNN	Stacking	Progressive Mechanism	ICD Hierarchy
MD-BERT [209]	Hierarchical BERT	Stacking	Label Attention	ICD Description
MSMN [204]	LSTM	Stacking	Multi-synonyms Attention	UMLS

Table 3. A summary of representative models under the unified encoder-decoder framework

3.1.1 Recurrent Neural Encoders. Recurrent neural networks model the temporal sequences via their internal states and capture sequential dependencies. Thus, they have been widely applied to textual sequence modeling and clinical note encoding. Generally, the recurrent neural encoder (in Fig. 4a) outputs a hidden representation $\mathbf{H}^l \in \mathbb{R}^{n \times d_h}$ of the l-th layer:

$$\mathbf{H}^l = \text{RNN}(\mathbf{X}), \tag{1}$$

where n is the number of words and d_h is the dimension of the hidden representation. However, the vanilla RNN-based model suffers from the vanishing gradient issue [70]. Shi et al. [157], one of the first works on applying RNNs for medical coding, developed an Attentive LSTM network. This model encodes clinical descriptions and long titles of ICD codes jointly with hierarchical text representations and uses an attention mechanism for matching important diagnosis snippets. Catling et al. [29] compared the TF-IDF (term frequency-inverse document frequency) feature with the word embedding features learned with the simplified gated recurrent unit (GRU). Mullenbach et al. [131] used a GRU with bi-direction as a baseline system for medical coding, where the last hidden representations are used for classification. From their pilot experiments, GRU shows more robust predictive performance than the LSTM network-based coding model. Blanco et al. [15] studied capabilities of various RNN models such as GRU and ELMo (Embeddings from Language Model). Other follow-up works such as HA-GRU [11] and HLAN [50] further improved the vanilla BiGRU with hierarchical attention, including

two levels on sentence and document representations. Hierarchical attention can help mitigate the difficulty of encoding long text sequences. Sec. 3.1.5 introduces more details of hierarchical encoders.

3.1.2 Convolutional Neural Encoders. The success of convolutional neural networks in computer vision inspires researchers to use convolutional architecture for medical coding. The TextCNN model [96] acts as a simple but essential baseline. The convolutional layer extracts local features from pretrained or randomly initialized word vectors. Fig. 4b illustrates the CNN-based text encoder. The representation with max-pooling is then used for medical code classification. Karimi et al. [87] compared standard CNN architecture with conventional classifiers such as decision trees and support vector machines [59, 163] on both in-domain and out-of-domain data and showed that CNN architectures with optimal parameter settings gain comparable results with conventional methods on sparse and skewed data. CAML [131] combines multiple-filter CNN-based text encoders and an attention decoder (introduced in Sec. 3.3). DCAN [77] develops dilated convolution layers, which apply convolutions with dilated filters to increase the receptive field. Given a sequence of one-dimensional elements $\mathbf{x} \in \mathbb{R}^n$ and a convolutional filter $f : \{0, \ldots, k-1\} \rightarrow \mathbb{R}$, the hidden representation in the l-th layer of stacked dilated convolution layers is calculated as $\mathbf{H}^l_{ij} = \sum_{j=0}^{n-1} f(j) \cdot \mathbf{x}_{s-d_l \cdot j}$,

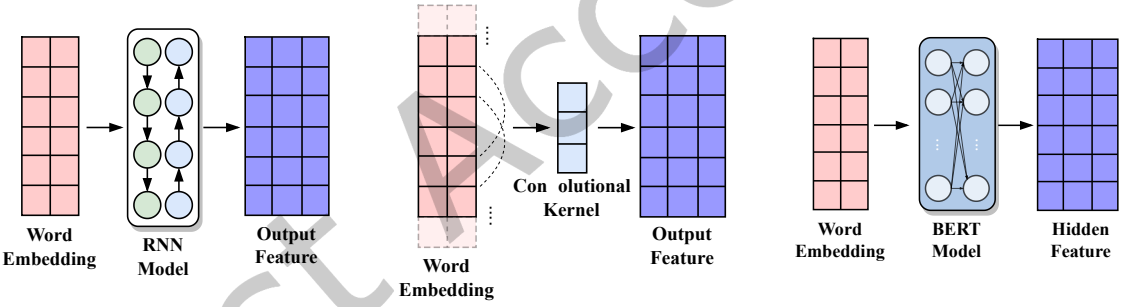
$$\mathbf{H}^l_{ij} = (\mathbf{w}_i *_{d_l} f)(\mathbf{w}_{ij}) = \sum_{j=0}^{n-1} f(j) \cdot \mathbf{x}_{s-d_l \cdot j}, \tag{2}$$

where d_l is the dilation size of the spacing between kernel elements in the l-th layer, s is the element of input sequence, and $s - d_l \cdot i$ refers to past time steps. When stacking a deeper architecture, the dilation size is exponentially increased to expand the receptive field. Other models also use the CNN-based text encoder. For example, MultiResCNN [104] concatenates the features of multi-filter convolutions. Similarly, MVC-LDA [153] introduces multi-view CNN by applying max-pooling over different channels with different convolutional filters. Ji et al. [82] developed a Gated CNN encoder that uses an LSTM-style gating mechanism to control the information flow. The Fusion model [123] deploys a Compressed CNN module that applies attention-based soft-pooling over word convolution features, reducing the number of word representations. Critical entities can help to recognize the correct medical code. ECNN [31] enhances the CNN model with entities extracted from the input text. Inspired by the squeeze-and-excitation network [73], EffectiveCAN [117] stacks multiple residual squeeze-and-excitation blocks with convolutional operations.

3.1.3 Neural Attention and Transformer Encoders. The neural attention mechanism computes a weighted sum of vector values of hidden representations dependent on the query vectors. Compared to RNN and CNN, self-attention has been widely adopted for transfer learning, i.e., as building blocks for large pre-trained language models. This allows leveraging the linguistic associations from massive corpora for subsequent tasks. The superior performance gained by BERT attracts researchers of medical coding to apply BERT-based text encoders, as shown in Fig. 4c. However, due to the complexity of the self-attention mechanism, only a few works use pure attention-based encoders to model the clinical notes especially discharge summaries. Since 2021, more researchers have proposed to use transformer-based models. TransICD [14] applies transformer text encoder and structured self-attention to learn representations. Coutinho et al. [40] used Transformers for ICD-10 coding from Portuguese text. Some attempts explore the possibility of BERT encoders. Roitero et al. [152] built a BERT model via domain-specific pretraining and fine-tuning. BERT encoders are limited to encoder the maximum sequence length of 512. When dealing with long documents, they do not achieve superior performance compared with CNN or RNN-based encoders, potentially due to the limitation of BERT to encode long documents and keywords according to Gao et al. [63]. Thus, BERT encoders are usually used to encode the long clinical notes in a hierarchical manner, which will be introduced in Sec. 3.1.5.

More recent studies attempt to study the performance of efficient transformer-based methods. For example, Feucht et al. [61] found that Longformer achieves better results than BERT. Yogarajan et al. [201] applied concatenated representations from contextualized language models and used Longformer [12] and Transformer-XL [44] to processed longer sequences. Yang et al. [200] adopted longformer with domain-specific knowledge enhancement. Hou et al. [72] integrated the long-distance dependency features captured through Clinical-Longformer with code synonyms, code hierarchy, and code co-occurrence knowledge to improve long-tail classification. Gomes et al. [67] compared the ability of the chunk encoder and longformer encoder for lengthy text modeling. Michalopoulos et al. [127] applied BigBird [205] designed for long sequence encoding to encode discharge summaries. Niu et al. [135] used the FLASH [74], a variant of Transformer, as a feature extractor to extract meaningful semantic features from long clinical notes. Liu et al. [115] pre-trained the new language model ClinicalplusXLNet based on fine-tuning the pre-trained Transformer model. The authors conducted continuous pre-training using clinical corpus from MIMIC-III using XLNet-Base. Subsequently, Duan et al. [53] employed ClinicalplusXLNet as the encoder, encoding the segmented clinical text to obtain semantic features. Xie et al. [192] developed a knowledge-based dynamic prompt learning algorithm for coding prediction. The method utilizes various masked language models and dynamically generates prompts based on personal medical information and medical knowledge graphs to provide valuable information representation for the model training.

Previous methods rely on existing pretrained language models to obtain contextualized embeddings. Zhang et al. [211] proposed BERT-XML that combines BERT encoders with multi-label attention. Rather than fine-tuning the pretrained BERT encoder, the authors trained the self-supervised BERT-XML encoder from scratch on clinical notes to solve the out-of-vocabulary issue. Moreover, they pretrained the BERT-XML model with a sequence length of 1024 for long sequences.

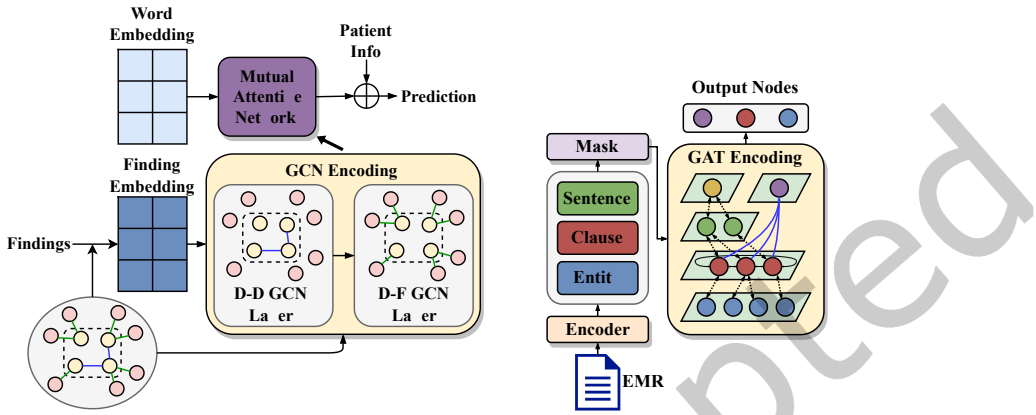


(a) RNN encoder (b) CNN encoder (c) BERT encoder

Fig. 4. Illustrations of representative neural text encoders. (a) The RNN encoder captures sequential dependency. (b) The CNN encoder extracts local features. (c) The BERT encoder encodes contextualized information.

3.1.4 Graph Encoders. Many natural language processing tasks construct text graphs and adopt graph neural networks as text encoders to learn textual features [190]. Several works in medical coding also use graph-based encoders to capture the structural information during the diagnosis process. Yuan et al. [203] built a medical graph that consists of diseases and findings and deployed the Graph Convolutional Network (GCN) [98] to learn graph representations. The authors considered disease-disease (D-D) and disease-finding (D-F) graphs as shown in Fig. 5a during the encoding of disease hierarchy and causal relations. Similarly, Lu et al. [121] transformed text features extracted by pre-trained BERT into node representations in heterogeneous graphs and utilized GCN for message passing. CMGE [188], a multi-granularity graph-based method, builds a hierarchical

graph that contains four types of nodes: general nodes (patients' age and gender), sentence nodes, clause nodes, and entity nodes, as illustrated in Fig. 5b. It uses the Graph Attention Network (GAT) [171] for information aggregation. The multi-granularity graph reasoning enables supporting fact extraction from the clinical notes and explainable diagnosis prediction. Luo et al. [122] constructed a code relation graph to capture the complex interaction relationships between ICD codes and improve code allocation accuracy.

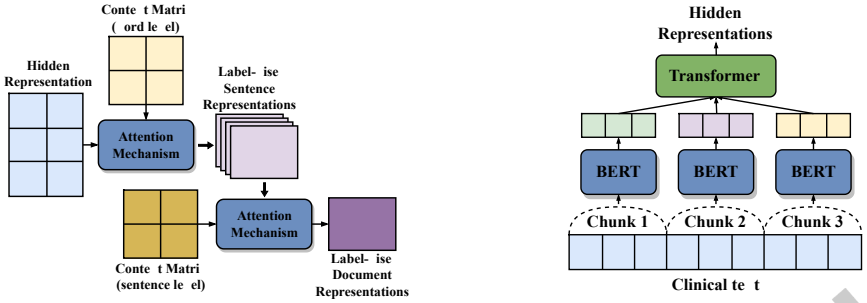


(a) GMAN [203] (b) CMGE [188]

Fig. 5. Graph encoders on the constructed medical graph. a) GMAN applies the GCN layers to encode disease hierarchy and disease-finding causal relations. b) CMGE uses the GAT as the graph encoder on a multi-granularity graph.

3.1.5 Hierarchical Encoders. Several methods adopt hierarchical text encoders, as a "meta"-encoder with the above encoding modules, to take the hierarchical structure of documents into encoding and potentially solve the difficulty in encoding lengthy clinical documents that encode hierarchical elements of the long documents such as characters, words, sentences, and chunks. Shi et al. [157] built a hierarchical encoder with character representation, word representation, and sentence representation. Dong et al. [50] adapted hierarchical attention networks with label-wise word-level and sentence-level representations for an improved attention-based explanation for each code (in Fig. 6a). To make BERT-based text encoders compatible with long clinical notes, Ji et al. [78] developed BERT-hier that divides long notes into chunks and uses another Transformer network to encode the embeddings of different chunks (in Fig. 6b). Although the hierarchical BERT-based encoder improves the performance, it is still not as good as advanced CNN or RNN-based models. Pascual et al. [142] conducted a similar study. A recent work called Medical Document BERT (MD-BERT) [209] proposes a more advanced hierarchical encoding method by considering token-level, sentence-level, and document-level representation learning and attaching the classification layer to any levels of interest according to specific tasks. This model achieves better performance than previous attempts on utilizing transformers-based text encoders. Other recent findings also show that BERT-based encoders can achieve improved performance with better configuration and training when handling long texts. For example, Dai et al. [43] showed that the document splitting strategy for text encoders is important. Afkanpour et al. [2] found that the utilization of token-level representation and longer text sequence can improve performance. In addition to discharge summaries, text metadata such as time and note type is also used in the hierarchical transformer model to improve temporal document sequence encoding [133].

3.1.6 Summary. Neural encoders play an important role in learning rich representations from clinical notes. Early research on deep learning-based medical coding explored recurrent and convolutional neural networks and



(a) Hierarchical encoding [50]

(b) Hierarchical BERT [78]

Fig. 6. Illustrations of the hierarchical encoder and decoder. (a) The hierarchical encoder learns hierarchical representations of words, sentences, and documents. (b) The hierarchical BERT encoder processes text chunk by chunk and aggregates the embeddings via an additional Transformer.

achieved improved performance than feature engineering-based methods. Many follow-up works further improve CNNs and RNNs to enhance their capacity to capture long context. Self-attention-based Transformer networks suffer from quadradic complexity. However, recent studies on hierarchical encoders and efficient transformers are getting a better performance for medical coding. Graph neural networks that can capture structural information represented in heterogeneous text graphs are emerging. The inductive bias in different neural encoders is the key consideration for the choice of encoders. However, there is no clear evidence on which neural encoder is optimal. One recommendation is to choose the neural encoder based on the data and the need for coding practice. Neural architectures offer benefits such as automated feature learning, representation hierarchies, and performance improvement in medical coding tasks. However, their black-box nature, model complexity, lack of intuitive representations, and the need for additional explainability techniques pose significant challenges in achieving interpretability and explainability. Future research requires to develop methods to address these limitations and make neural architectures more transparent and understandable for medical coding applications.

3.2 Building Deep Architectures

Most existing neural network-based medical coding models have deep architectures. The most straightforward approach uses stacking to build deep neural architectures, such as stacking multiple recurrent layers and hierarchical components of different levels of elements as in multi-layer perceptrons. Also, different neural blocks can be stacked into deep networks, for example, the recalibrated aggregation module [161] with multiple convolutional layers is built upon a bidirectional GRU network, the MSATT-KG [193] stacks densely connected convolutional layers and multi-scale feature attention, and the BiCapsNetLE [10] deploys a capsule neural network upon the BiLSTM layer to extract features further.

When encoding long clinical notes with very deep architectures, features learned by higher layers tend to capture abstract features but sometimes miss some vital information. Ji et al. [82] proposed to use embedding injection to mitigate the information loss with the increase of neural layers. The embedding injection concatenates the original word embeddings into each intermediate layer of the backbone network as:

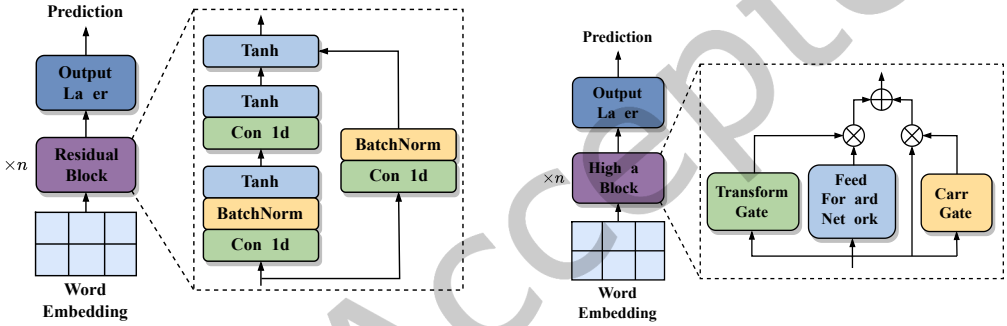
$$\mathbf{J}^l = \text{concat}\left[\mathbf{X},\ \mathbf{H}^l\right], \tag{3}$$

where $\mathbf{J}^l \in \mathbb{R}^{n\times(d_e+d_h)}$ are the features with original embeddings injected.

The most widely used approach to building deep networks for automated medical coding is to use residual connections, as shown in Fig. 7a. Deep residual learning introduces the skip connection to avoid the effect of the vanishing gradient. It enables the building of very deep neural network architectures. Given the input encoding vector $\mathbf{x}$, the output of residual connection is denoted as $o = \sigma(\mathbf{x} + \mathcal{G}(\mathbf{x}))$, where $\mathcal{G}$ represents neural layers and σ is a non-linear activation function. Several medical coding models use residual networks between stacked layers, which are denoted as:

$$\mathbf{H}^{l+1} = \sigma(\mathbf{H}^{l} + \mathcal{G}(\mathbf{H}^{l})). \tag{4}$$

MultiResCNN [104] is the first to combine residual learning with the concatenation of multiple channels with different convolutional filters. Other follow-up works such as DCAN [77] and Fusion [123] also use the residual neural network. We also illustrate the highway networks for building deep architectures in Fig. 7b, although no existing medical coding models adopt the highway mechanism. Highway networks use the gating mechanism (i.e., the transform gate and the carry gate) to control the amount of input information and avoid attenuation when stacking very deep layers. The highway networks can be an alternative to building deep medical coding models.



(a) Residual networks (b) Highway networks

Fig. 7. Illustrations of residual networks and highway networks for building deep architectures

3.3 Decoder Modules

After the encoder modules have extracted hidden representations of clinical notes, the decoder modules map the learned representations into medical codes as the final classification results via a decoding process. The hierarchical and large-scale characteristics of medical codes have promoted the design of various decoder modules. This section introduces four main types of decoder modules, including the fully connected layer-based decoder that offers simplicity and efficiency (Sec. 3.3.1), neural attention decoders that enhance code prediction by focusing on relevant information (Sec. 3.3.2), hierarchical decoders that leverage the hierarchical structure of medical codes for more structured predictions (Sec. 3.3.3), multitask decoders that handle multiple coding systems simultaneously (Sec. 3.3.4), and few-shot decoders aim to solve the few-shot learning problem and make accurate predictions with minimal examples, making them valuable in challenging coding situations (Sec. 3.3.5).

In addition to these decoders, pipeline-based methods further deploy some post-hoc modules to boost performance. For example, Tsai et al. [169] proposed a two-stage method, i.e., a candidate generation stage to generate candidate sets of ICD codes and a candidate reranking stage that leverages the label correlation to rerank the generated code sets. Some non-parametric post-processing methods can also be applied for adjusting the decoders

of medical coding models; for example, the Classification with Alternating Normalization method [83] that redistributes the prediction probability.

3.3.1 Fully Connected Layer. The most straightforward decoder module is a linear fully-connected layer, widely used in many classification tasks. The prediction logits $\hat{y} \in \mathbb{R}^m$ between 0 and 1 are produced by the Sigmoid activation function with a pooling operation over the linearly projected matrix, calculated as:

$$\hat{y} = \text{Sigmoid}(\text{Pooling}(\mathbf{H}\mathbf{W}^{\text{T}}), \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{m \times d_h}$ are the linear weights for m medical codes. Medical coding models that use a linear layer as a decoder include Attentive LSTM [157] and C-MemNN [147].

3.3.2 Neural Attention Decoders. The neural attention mechanism has also been applied to decoding in addition to its usage for encoding clinical notes introduced in Sec. 3.1.3. One useful attention mechanism for decoding is the so-called Label-wise Attention Network (LAN) which prioritizes important information in the hidden representation relevant to medical codes. The LAN-based decoder, as illustrated in Fig. 8a, uses the dot product attention to calculate the attention score $\mathbf{A} \in \mathbb{R}^{n \times m}$ as:

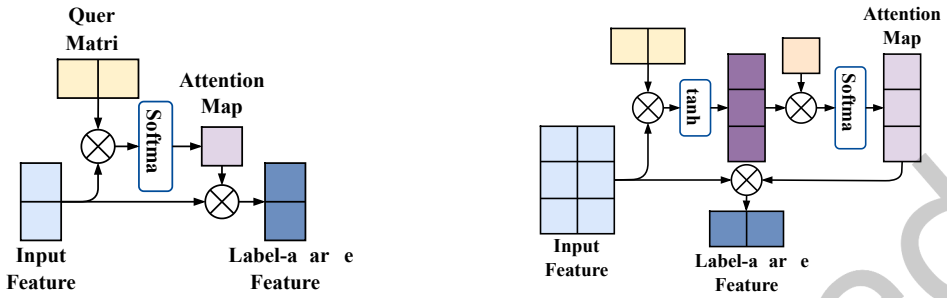
$$\mathbf{A} = \text{Softmax}(\mathbf{H}\mathbf{U}), \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{h_L \times m}$ is the query matrix of the label attention layer for m medical codes, and h_L is the dimension of the query. By multiplying attention $\mathbf{A}$ with the hidden representation, i.e., $\mathbf{A}^{\text{T}}\mathbf{H}$, the output of the attention layer is obtained for medical code prediction. CAML [131] is the first to apply LAN by using the attention matrix to capture the importance of ICD code and hidden word representation pair. DCAN [77] and MultiResCNN [104] also use the LAN decoder as a building block of their models. Fusion [123] deploys similar code-wise attention after feature aggregation and RAC [93] implements the code-title guided attention module. The LAN-based decoder preserves sequential information captured by the text encoder and enables label awareness to benefit medical code classification. JLAN [106] proposes a dual attention mechanism that combines self-attention and label attention. LAAT [173] applies the structured self-attention [110] (in Fig. 8b) that projected the hidden representation via a linear transformation and non-linear activation as:

$$\mathbf{H}' = \tanh(\mathbf{W}_s\mathbf{H}), \tag{7}$$

where $\mathbf{W}_s \in \mathbb{R}^{h \times d_h}$ is a weight matrix, and h is the number of hops of the structured self-attention. In practice, LAAT sets the number of attention hops to the number of labels. Similarly, TransICD [14] uses the structured self-attention mechanism to achieve code-specific decoding for automated medical code prediction. Attention-based decoders enhance medical code prediction by modeling code information. However, due to the lack of training data, code interaction cannot be effectively learned, especially for those rare codes. Zhou et al. [214] proposed an interactive shared representation network to enhance the interaction among code-relevant information via multi-layer transformer decoders. To capture code co-occurrence, the authors further implemented two additional tasks, i.e., missing code completion and wrong code removal. Wu et al. [191] designed a joint attention decoder that utilizes document-based attention to extract text information and label-based attention to emphasize the semantic connection between label semantics and document content. To balance the contribution of document-based attention and label-based attention to label feature representation, the authors employed layer and gate mechanisms to achieve adaptive fusion. Due to the existing label attention mechanism to identify critical segments in the entire text at once, it may ignore some crucial local information scattered in paragraphs. Kim et al. [94] designed a new neural decoder composed of two label attention layers by integrating traditional and partition-based label attention mechanisms to obtain global and local potential feature representations. Partition-based label attention divides the text representation obtained from the encoder and generates label-specific features for each segment. Then, the weighted summation of features is performed to obtain a combined label-specific feature

matrix. Inspired by the coding process of clinical coders (selecting general categories first and then specific subcategories), Nguyen et al. [134] designed a two-stage decoding process that utilizes attention mechanisms first to predict the parent code and then the child code based on previous predictions.



(a) Label attention (b) Structured self-attention

Fig. 8. Illustrations of attention encoders. (a) Label attention learns label-aware representations for decoding. (b) Structured self-attention also learns label-specific representations.

3.3.3 Hierarchical Decoders. Hierarchical models that use the hierarchical code structure have been studied to improve automatic coding a long time ago [45]. Building hierarchical decoders is still a promising research direction in the recent advances in deep learning-based methods. JointLAAT [173] proposes a hierarchical joint learning method that produces the code prediction level by level according to the ICD hierarchy. Firstly, the model predicts the normalized ICD codes with the first three characters. Then, the first level's predictions are projected back to a vector and concatenated with the label-specific representation of the second level in the ICD hierarchy for the final prediction. An earlier work by Falis et al. [58] uses three hierarchical decoding layers for ICD codes, as shown in Fig. 9. The hierarchical decoding-based JointLAAT outperforms the vanilla LAAT slightly in some evaluation metrics. There is still room for improvement by making use of the hierarchical nature of the medical coding system. Unlike those hierarchical decoding methods via joint learning, RPGNet [180] formulates the medical coding task as a path generation problem and proposes a coarse-to-fine ICD path generation model based on adversarial reinforcement learning. It leverages a path generator to generate paths and a path discriminator to distinguish the generated paths from positive paths. Liu et al. [114] proposed hierarchical label-wise attention in response to the hierarchical encoding at token and chunk levels.

3.3.4 Multitask Decoders. Multitask decoders predict medical codes with multiple task branches powered by multitask learning [28]. Tsai et al. [168] took low-level code and high-level category as two task branches in their multitask learning framework. Medical coding models aforementioned in this section build decoders for a single coding system. However, several different systems have been used for different purposes. To enable decoding of multiple coding systems and utilize the joint learning of similar tasks, MT-RAM [161] deploys a multitask decoding scheme that includes two branches with label-wise attention for ICD and CCS code prediction as shown in Fig. 10a. As a following-up work, MARN [162] improves the multitask decoders with the focal loss to balance the learning of codes with imbalanced code frequencies. More publications introduce other auxiliary tasks to train joint learning models. Wiegreffe et al. [185] predicted the outputs of the Apache clinical Text Analysis Knowledge Extraction System (cTAKES)[3] together with ICD codes as illustrated in Fig. 10b. CMGE [188]

[3]Available at https://ctakes.apache.org/

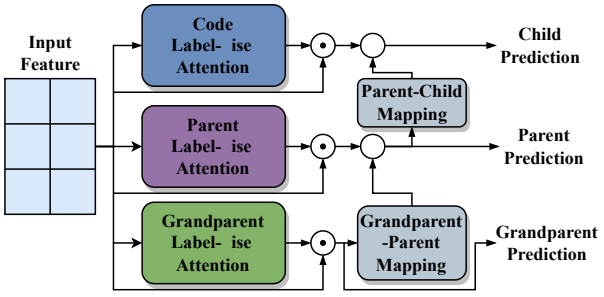
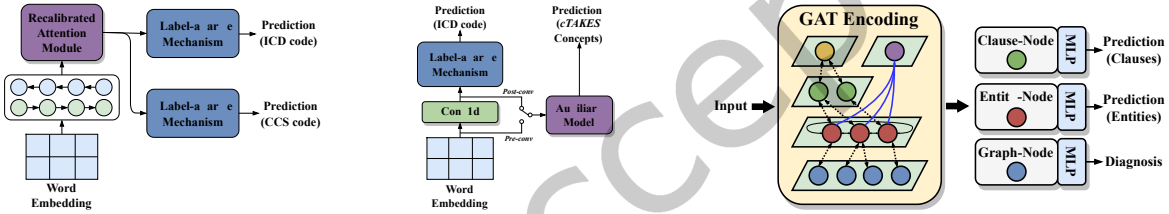


Fig. 9. Illustrations of the hierarchical decoder that considers the hierarchical structure of medical ontology during code prediction [58]. Circles with dots represent dot production and circles with "C" denotes concatenation.

in Fig. 10c considers graph classification, sub-sentence classification, and entity classification. Rios et al. [150] jointly trained a multitask learning model with losses for topography and histology codes and the hierarchical loss with a hierarchical regularization.



(a) MT-RAM [161] (b) Wiegreffe et al. [185] (c) CMGE [188]

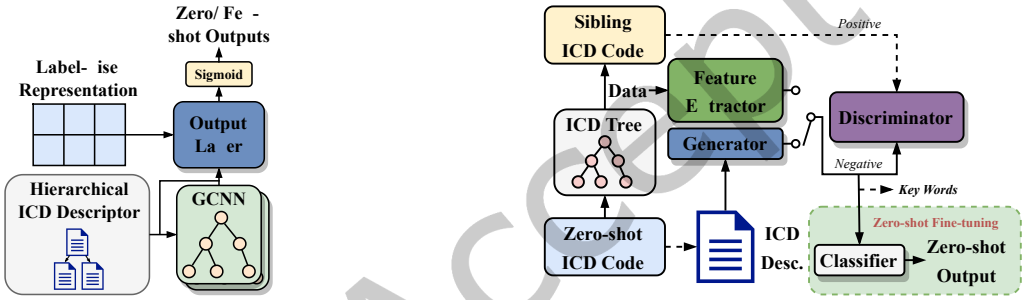
Fig. 10. Illustrations of multitask decoders. (a) MT-RAM adopts two-branch joint multitask training; (b) Wiegreffe et al. used distantly surprised cTAKES output prediction as a task head; (c) CMGE adopts three classification tasks for graphs, sub-sentences, and entities.

3.3.5 Few-shot/Zero-shot Decoders. The automatic medical coding task has a large label space, with some frequently appearing and many labels never shown in the dataset. Few-shot models aim to predict codes that only appear a few times in the training data, and zero-shot models aim to predict codes that never appear in the training data. Rios and Kavuluru [151] are among the first work of medical coding in the few-shot and zero-shot settings. They defined the few-shot and zero-shot coding problem as a retrieval task (in Fig. 11a) in which the model calculates the code probability as the semantic matching between the representations of clinical documents and label vectors of target codes, denoted as:

$$\tilde{y}_i = \text{Sigmoid}(\mathbf{e}_i^\top \mathbf{v}_i), \tag{8}$$

where $\mathbf{e}_i$ is the label-specific document vector and $\mathbf{v}_i$ is the label vector for i-th label. The proposed model ZAGCNN uses CNN layers to extract features of clinical notes and GCN layers to encode the ICD code hierarchy boosted with code descriptions. Following the similar few-shot and zero-shot setting, Lu et al. [120] improved the ZAGCNN model with knowledge aggregation from multiple graphs, i.e., the predefined hierarchy, the semantic similarity graph of label description, and the label co-occurrence graph. Meta-LMTC [177] extends the ZAGCNN model by using optimization-based Model-Agnostic Meta-Learning (MAML) algorithm [62] and

two sampling strategies (i.e., instance- and label-based) for meta optimization. Unlike the ZAGCNN and its extensions, Song et al. [159] adopted the generalized zero-shot learning method for ICD coding. The authors proposed an Adversarial Generative Model (AGM) and utilized the Hierarchical Tree (HT) and code descriptions to generate code-specific features with the generative adversarial networks, as shown in Fig. 11b. The generated features are further used to fine-tune the medical coding model for zero-shot codes. Those methods mentioned above rely on external knowledge sources such as code hierarchy and descriptions (see more introduction about the usage of auxiliary information in Sec. 3.4, specifically Sec. 3.4.2 for code descriptions and Sec. 3.4.3 for code hierarchy). CoGraph [179] constructs a heterogeneous word-entity graph to represent clinical notes and performs graph contrastive learning on the constructed graph to improve the model's capability on few-shot prediction. Contrastive learning explores the intra-correlation of word-entity graphs via sampling and the inter-correlation of word-entity graphs via sequential modeling of graphs at different clinical stages. During the graph construction of the CoGraph model, Wikipedia acts as the source of external knowledge to obtain entity nodes. Ji et al. [80] showed that task-conditioned parameter generation with additional task information improves zero-shot diagnosis prediction. Auxiliary knowledge plays an essential role in few-shot and zero-shot medical coding.



(a) ZAGCNN [151] (b) AGM-HT [159]

Fig. 11. Illustrations of few-shot and zero-shot medical coding. (a) ZAGCNN considers few-shot/zero-shot medical coding as a retrieval problem. (b) AGM-HT utilizes adversarial generative training.

3.3.6 Autoregressive Generative Decoders. A recent pertaining and fine-tuning paradigm has been used for medical coding as introduced in Section 3.1.3. However, in many cases, there is a significant gap between the goals of the downstream tasks and the pretraining goals. Moreover, specific fields require large amounts of the supervised corpus during fine-tuning. A new fine-tuning paradigm based on the pretrained language model, prompt tuning that gives some best cues as the task-specific context for pretrained generative language models, has emerged to address these challenges. This approach has proven effective in few-shot tasks [64, 103]. In medical coding, Yang et al. [200] addressed the long-tail challenge using a prompt-tuning technique for label semantics, representing the first attempt to apply prompts to multiple label classification tasks. Specifically, the authors added a series of ICD code descriptions as the prompt and incorporated them early with clinical notes. To further improve the performance of the medical coding, the authors proposed a knowledge-enhanced longformer that injected three domain-specific knowledge (hierarchy, synonyms, and abbreviations) and utilized comparative learning for additional pre-training. In a follow-up study, Yang et al. [199] further tackled the long tail challenge in multi-label classification by converting it into an autoregressive generation task. The authors exploited a SOAP structure (i.e., subjective, objective, assessment, and plan) to generate free text diagnoses and procedures, which is medical

logic used by physicians to note clinical documentation. They then translated the generated text's description to infer ICD codes with clinical vocabulary constraints, which solves the hallucination issues of generative models. Prompt tuning with generative language models provides a novel solution for medical coding. It has benefits to utilize the knowledge from large language models. However, it also has some limitations. For example, the hallucinated generation can dampen the coding accuracy and require some engineering efforts on controlled generation, specifically when there is a shift in coding guidelines. Also, autoregressive generative models have the scaling issue when generating tokens, making them slower than non-autoregressive methods [199].

3.3.7 Summary. The problem setup and the principle of learning paradigms are the main motivators for choosing the decoder module. Neural attention decoders improve the fully connected layer-based decoder in the standard supervised learning setup to prioritize the representation learning on important information. The hierarchical decoders fit the hierarchical nature of medical codes. Multitask decoders aim to predict medical codes of multiple coding systems. Few-shot and zero-shot decoders solve the learning problem with rare or unseen medical codes. And the autoregressive generative decoders, as an emerging approach, utilize the reasoning capacity of large autoregressive language models. The choice of decoder module in medical coding models should align with the problem's nuances and learning paradigms. Each type of decoder offers unique advantages, from enhanced attention mechanisms for information prioritization to specialized hierarchical structures, multitasking capabilities, adaptability to rare codes, and advanced autoregressive reasoning. Careful consideration of these factors is essential to develop effective and context-aware medical coding systems that can meet the diverse needs of healthcare professionals and patients.

3.4 Usage of Auxiliary Information

Auxiliary information can be utilized to enhance representation learning and improve the performance of medical coding. This section introduces the usage of auxiliary information, including implicit information such as label information via randomly initialized embeddings and explicit information (or external data) such as Wikipedia articles, textual code descriptions, and code hierarchies. Implicit label information has been used by most previously introduced label attention-based models. The joint embedding model (LEAM) [175] embeds labels and leverages the compatibility between word and label embeddings to calculate attention scores. The following paragraphs review the methods that use external data explicitly. The external data can be applied to both encoders and decoders. When applied to encoders, external data enhance the representation learning of clinical texts. The external information usually acts as the regularization for decoders when combining external data augmentation with the decoding process.

In addition to explicit usage of auxiliary information, data augmentation methods can also be applied to enrich the training data. Kim and Ganapathi [93] introduced a simple sentence permutation method to augment the training data three times and improve code prediction performance.

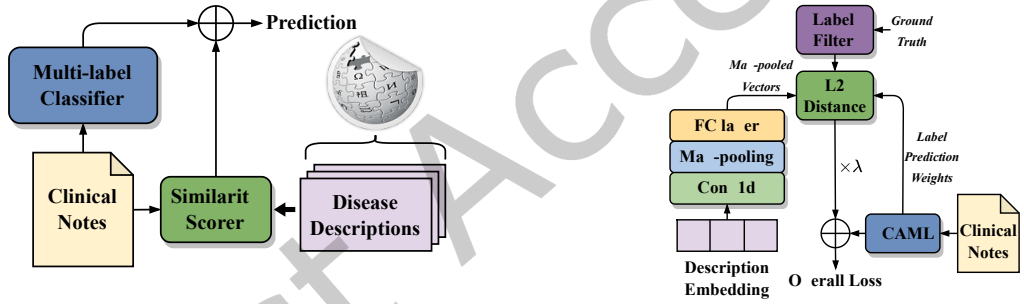
3.4.1 Wikipedia Articles. Wikipedia articles explain medical diagnoses in detail and are used to enhance the deep learning model on clinical text understanding. Prakash et al. [147] resorted to Wikipedia as an external knowledge source. Specifically, the authors used term search to find relevant articles to the diagnoses in clinical notes. They proposed C-MemNN with an iterative condensation of memory representations that utilize external knowledge sources from Wikipedia to enhance memory networks by preserving the hierarchical structure in the memory. KSI [9] in Fig. 12a uses element-wise multiplication and attention mechanism to fuse the knowledge from Wikipedia articles into clinical notes. There are 389 available Wikipedia pages when considering the first three digits of ICD-9 diagnosis codes. The KSI model defines the medical coding task as a classification problem of 344 ICD codes found in the code vocabulary of the used dataset. Following the same setup, MCDA [183], a medical concept-driven attention model, aligns the clinical notes and Wikipedia articles in the latent topic

space based on topic modeling. The joint embedding or alignment of Wikipedia and clinical notes introduces external knowledge sources to medical coding models. However, because some specific medical codes have no corresponding Wikipedia pages, the usage of KSI is only limited to coding three-digit ICD-9 codes, i.e., diagnostic category classification. The absence of fine-grained coding may lead to the ineffectiveness of medical coding models in rare diagnoses or procedures.

3.4.2 Code Description. The textual description of medical codes describes the exact meaning of codes and provides extra semantic information for abstract codes. The embeddings of code description are denoted as $\mathbf{D} \in \mathbb{R}^{m \times d_t}$, where m is the number of codes, and d_t is the dimension of description embedding. Several publications utilize the code description to enhance representation learning. DR-CAML [131] as shown in Fig. 12b uses the word vectors of description as a regularization when optimizing the label-wise attention module. Similarly, CAIC [165] develops cross-textual attention to establish the connection between medical notes and ICD codes. GatedCNN-NCI [82] builds fully connected interaction between notes and codes. BiCapsNetLE [10] uses embeddings of ICD descriptions to inject label information into the word embeddings of clinical notes and the features learned by capsule networks. DLAC [61] proposes a description-based label attention that computes the label attention matrix with the description matrix and transformed hidden representation matrix as

$$\mathbf{A} = \text{Softmax}\left(\mathbf{HU} \cdot \mathbf{D}^{\top}\right), \tag{9}$$

where $\mathbf{U} \in \mathbb{R}^{d_h \times d_t}$ is a transformation matrix that aligns the dimensions of the hidden representation and the description matrix. A prompt-based fine-tuning model [200] adds a series of ICD code descriptions as the prompt to integrate code description and input notes for multi-label few-shot ICD coding.

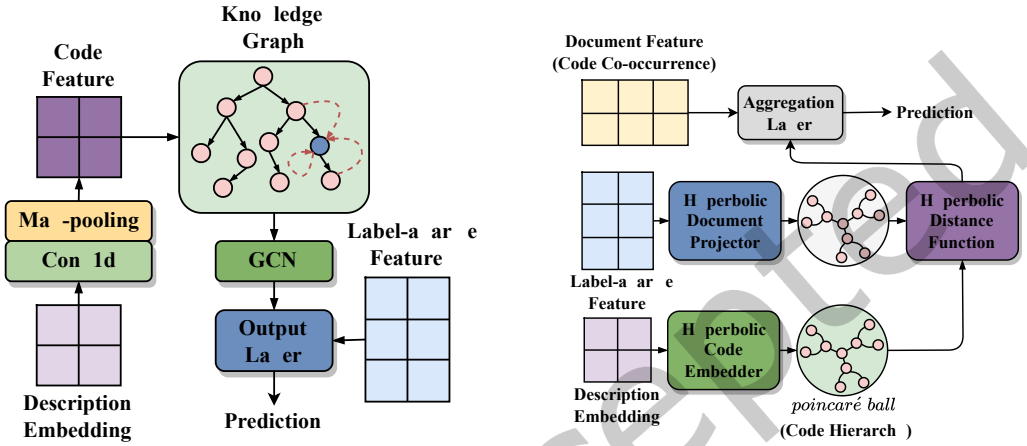


(a) KSI [9] (b) DR-CAML [131]

Fig. 12. Illustrations of models that DR-CAML and KSI infuse external text features for regularization and feature augmentation. (a) KSI augments features via the multiplicative interaction between note features and embeddings of Wikipedia articles (b) DR-CAML infuses code description via regularization.

3.4.3 Code Hierarchy. As introduced in Sec. 3.3.3, hierarchical decoders use the code hierarchy. MSATT-KG [193] in Fig. 13a infuses code hierarchy into document representation via structured knowledge graph (KG) propagation and label-dependent attention, where the code hierarchy is treated as a KG, and the graph convolutional network (GCN) is used to capture code relationships. Similar to MSATT-KG, HyperCore [26] also uses GCN to encode the code co-occurrence. Besides, it utilizes hyperbolic embedding and co-graph representation with code hierarchy, as shown in Fig. 13b. HieNet [181] builds a bidirectional hierarchy passage encoder, consisting of a bidirectional passage retriever and a tree position encoder, to represent the code hierarchy with semantic and positional features. When classifying frequent codes with no significant hierarchical connections, Michalopoulos et al. [127] built a

co-occurrence graph of ICD codes with edge weights measured by normalized point-wise mutual information and applied graph convolutional networks to encode the ICD codes. Lu et al. [119] represented the ICD hierarchy as a super-tree and introduced tree editing distance [208] to capture disease relationships at the code hierarchy. The hierarchical structure of the code system is a unique characteristic of medical coding, especially for predicting the complete code set. It is an exciting research direction that has the potential to improve coding performance and produce reliable and interpretable coding results.



(a) MSATT-KG [193] (b) HyperCore [26]

Fig. 13. Illustrations of models that incorporate ICD hierarchy. (a) MSATT-KG uses GCN to encode the ICD code hierarchy. (b) HyperCore also uses GCN but in hyperbolic space.

3.4.4 Medical Ontology. We generally consider ontologies as TBox (i.e., concept-level knowledge that consists of logical statements constraining concepts) and ABox (data-level knowledge that consists of assertions over instances) [8]. A taxonomy (or hierarchy) can be reduced from the TBox of an ontology through a reasoning process called classification [8, p. 34][4]. Also, taxonomy has a DAG (directed acyclic graph) structure. Ontologies in the context of medical coding encompass a wide spectrum of structures, each with distinct characteristics and relationships; they are broadly referred to as terminologies or classification systems [39]. In fact, ICD has the form of an ontology.

Apart from that, the Unified Medical Language System (UMLS) is a comprehensive collection of dictionaries and ontologies of in-domain concepts [16, 111]. The UMLS is mainly based on three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon and Lexical Tools. Some studies explore the significance of ontology comprehension in designing effective deep-learning models for medical coding tasks. MSMN [204] extracts ICD code synonyms from the UMLS and proposes multiple synonym-matching networks to encode synonym information. Dong et al. [52] leveraged the UMLS as an intermediary dictionary to extend the annotation vocabulary of rare diseases for their identification from clinical notes. Falis et al. [56] used UMLS (and their concept matching to ICD) to augment training data for medical coding. The ability of a model to capture and utilize these intricate relationships can greatly impact its accuracy in code prediction. Consequently, understanding

[4]The classification as a Description Logic reasoning process [8] is different from its usage in machine learning.

the specific ontology used in a medical coding task becomes crucial in selecting, adapting, or developing deep learning models.

While some ontologies, like the ICD, primarily rely on hierarchical relations without directional relationships, others feature more complex and nuanced relationships, such as "causes" and "is-caused-by" disease relations. These variations in ontology structures can have profound implications for the applicability and effectiveness of the deep learning approaches. The structures of different ontologies, such as the organization of concepts, relationships, and attributes, vary significantly [68]. Approaches optimized for purely hierarchical taxonomies may require adaptation, for example, when dealing with ontologies that incorporate causal relationships. Assume we have ontology A and ontology B and consider the representation of the relationship "caused by" between diseases and risk factors. In ontology A, this relationship might be straightforward, while in ontology B, it could involve nested attributes and additional complexities. For example, in ontology B, the relationship might be represented as "Cardiovascular Disease" is "caused by" "Genetic Factors" AND "Environmental Factors". Deep learning models optimized for simpler relationships might struggle to capture the nuances of complex relationships in ontology B. The variability in how relationships are modeled affects the model's ability to understand and predict based on different ontological structures.

In the context of ontological representation, particularly within the Web Ontology Language (OWL) under the EL profile [174], we exemplify the SNOMED CT [138] that employs existential restrictions with attributes to articulate intricate relationships, such as those involving attributes like "Due to" (e.g., Cataract of the eye due to diabetes mellitus (disorder)) which represent causal relations. Some complex relations can be represented as triples, e.g., (A, Due to, B), but other complex relations require more expressive representation, for example, which involves conjunctions and nested existential restrictions [32, 48]. In this aforementioned example, "Cataract of eye due to diabetes mellitus" is equivalent to a conjunction of various concepts and attributes under a nested expression: `Cataract of eye due to diabetes mellitus` $\equiv$ `Disease` $\sqcap \exists$ `RoleGroup`. ($\exists$ `FindingSite`. `Structure of lens of eye` $\sqcap \exists$ `AssociatedMorphology`. `Abnormally opaque structure`) $\sqcap \exists$ `RoleGroup`. $\exists$ `DueTo`. `Diabetes mellitus`[5]. For complex relations that can be represented with triples, knowledge graph embeddings are essential [81]. Addressing more logically complex relations, the use of OWL, particularly EL++ embeddings, becomes imperative. For instance, OWL2vec* [33] and EL++ geometrical embeddings [102, 194] exemplify this approach. Notably, despite the wealth of research, there appears to be a gap in exploring the embedding of OWL ontologies, such as in SNOMED CT, within the context of clinical coding. This observation suggests a promising avenue for future studies in this domain.

3.4.5 Chart Data. Patients' chart data (or structured data) that record the physiological conditions of a patient can be used to enhance the performance of code assignments. Multimodal machine learning methods use text and chart data to predict medical code. Wang et al. [178] proposed a multi-label annotation model that inputs topic embeddings from patient notes and feature encodings from patients' chart data. The diagnosis code assignment module incorporates a disease correlation graph to capture the disease correlation. Xu et al. [195] used texts including discharge summaries, radiology reports, nursing notes, and tabular data such as admission, lab events, and prescriptions. The authors developed an ensemble learning method with text CNN applied to text representation learning and a decision tree applied to transformed numerical features. The experimental results show that effective modeling of multimodal data can improve the model's robustness and accuracy. Liu et al. [118] proposed a tree-enhanced multimodal attention network, TreeMAN, to capture the decisive information in structured medical data in EMRs. The authors first processed structured medical data into tabular data, then inputted tabular data into a trained decision tree to obtain tree-based features. Finally, the text representation and tree-based features are fused into a unified multimodal representation through an attention mechanism.

[5]Attributes like `DueTo` and `AssociatedMorphology` are in UpperCamelCase form. The expression is in SNOMED CT version 2024-02-01 for the concept ID 43959009. For more examples, see SNOMED CT browser at https://browser.ihtsdotools.org/.

3.4.6 Entities and Concepts. The text mentions about medical codes in clinical notes contain rich world knowledge. Apart from code descriptions and Wikipedia articles, several works also utilize entities and concepts that abstract the expressions in clinical notes. These methods usually use existing clinical ontologies such as the Unified Medical Language System (UMLS). Entity recognition and concept extraction aim to augment the text feature or provide additional supervision signals. Wiegreffe et al. [185] combined information extraction and medical coding by utilizing the cTAKES knowledge extraction system to extract concepts and reported a negative finding of document-level clinical coding. Falis et al. [56] extracted in-text UMLS entities (and the matched ICD entities) with SemEHR [189] and MedCAT [101] to augment new coded training data with synonym and sibling code replacement, and reported improved results s few-shot and zero-shot coding. Yuan et al. [204] obtained code synonyms by aligning concepts in the UMLS. Inspired by the multi-head attention mechanism [170], the authors proposed multiple synonyms matching networks that take code synonyms as queries to match clinical texts and improve code prediction. Yang et al. [200] injected three domain-specific knowledge from UMLS, i.e., hierarchy, synonyms, and abbreviations, into a knowledge-enhanced longformer that mitigates data sparsity and improves model performance. Li et al. [109] introduced medical knowledge from UMLS to construct entity-level text heterogeneous graphs. The usage of external knowledge improves individual notes' local context feature extraction. Ge et al. [66] utilized medical guideline knowledge from the Regional Documents of Old Dominion EMS Alliance [6] and medical datasets to construct a heterogeneous graph of medical entities. This graph captures entity relationships to compensate for data scarcity during model training.

3.5 Human-in-the-loop

In medical coding, the designed model and system need to meet the use scenarios of clinical coders, reduce the manual coding cost and enhance coding accuracy [107, 108]. ICD coding models with explanations can support decision-making better. In clinical practice, domain experts tend to trust the prediction results with reasonable explanations [187]. Thus, human-in-the-loop artificial intelligence that collaborates AI models with human beings becomes a useful paradigm to support the development of computer-assisted clinical coding (CAC). Various studies exploit human-computer interaction and are applied to augmented reality, brain-computer interface, and user customization. Combining human intelligence with deep learning models for medical coding requires much effort. For example, a successful human-computer interaction must empower the human coder to improve coding efficiency and accuracy. One quantitative study on human-computer interaction showed that the computer system should cater to the user's diverse needs while ensuring efficient, effective, and safe interaction [99]. This section considers human efforts as a form of auxiliary information in automated coding models. It reviews the literature on medical coding methods relevant to human-in-the-loop learning systems, such as general computer-assisted clinical coding, active learning for annotation, explainability, and human evaluation.

3.5.1 Computer-assisted Clinical Coding. CAC is a continuously developing technology that can improve the accuracy and quality of clinical coding and relieve the pressure on clinical coding personnel by assigning diagnostic and procedure codes from EHRs to automate clinical coding. Campbell et al. [23] reviewed and discussed CAC literature, and their findings indicated that CAC positively impacts coding quality and accuracy. Additionally, clinical coding personnel should view CAC as an opportunity rather than a threat. CAC transforms medical coding into a knowledge-based environment, and the current role of clinical coding professionals is transformed into clinical coding editors or analysts [130]. However, the clinical coding editor still has the ultimate responsibility. They can reject any inappropriate clinical coding suggestions by CAC software and send a consultation letter to clinicians to clarify ambiguous or contradictory documents [158]. The automated clinical coding workflow must still follow the clinical coding principles and specifications. Using machine learning methods or computer-assisted

[6] https://odemsa.net/

clinical coding to extract practical information from EHRs and assign medical codes is the actual demand of each medical health organization. Biomedical-named entity recognition and linking (NER+L) is committed to extracting concepts from texts in EHRs. Searle et al. [156] integrated MedCATTrainer with the biomedical NER+L model, which leverages active learning to improve the underlying NER+L model. Moreover, they provided researchers with configurable interfaces to define annotations specific to their research problems. This interface makes specific annotations for configurable use cases of previously identified and linked concepts.

3.5.2 Active Learning. Deep learning models need many annotation examples for training, and domain experts need a high cost to annotate these data. Recent studies investigated the adoption of human-in-loop learning. For example, active learning lets human annotators focus on the most informative data samples. Thus, the cost of manual labeling can be reduced. Ferreira et al. [60] employed the active learning method to select the sample with an enormous amount of information, significantly reducing manual annotation costs while maintaining the model's performance. Specifically, the authors studied two strategies for selecting samples: uncertainty metric and correlation. The uncertainty metric assesses how uncertain the model is for a particular instance, while the correlation measures the similarity between instances.

3.5.3 Explainability. Human-in-the-loop systems in healthcare aim to strike a balance between leveraging the capabilities of machine intelligence while maintaining the expertise and judgment of healthcare professionals. In such systems, healthcare providers can trust the correctness of the machine intelligence system's output [42], and the explainability allows them to understand the system's reasoning. Teng et al. [166] designed an explainable G_Coder model that uses two methods to verify its explainability. The first is to employ attention to extract keywords from clinical notes and display the correlation between medical labels and evidence. The second is to use doctors to judge the results of attention allocation quantitatively. Oberste et al. [137] established a framework for characterizing user knowledge and prior knowledge included in explanations by reviewing knowledge-informed machine learning in healthcare. The authors encourage future research to improve the explainability of the model while considering user background and experience to stimulate more trust in clinical settings.

3.5.4 Human Evaluation. Human evaluation is vital in assessing medical coding systems. Kim et al. [92] compared automated medical coding systems to human coders and introduced the Read, Attend, and Code (RAC) model, which performed on par with human coders. They hired two professional coders to assign ICD-9 codes to 508 patient discharge summaries, establishing a human coding baseline. The human coding baseline exceeded the machine learning baseline by a factor of 3.9 in micro-Jaccard similarity.

4 BENCHMARKING AND REAL-WORLD USAGE

This section introduces the data for benchmarking medical coding models and the evaluation metrics to evaluate the performance.

4.1 Data

The MIMIC database, including MIMIC-II [154] and MIMIC-III [85], is currently the most popular data source for the experimental study of medical coding. MIMIC-IV-Note [86] - the latest development of the MIMIC database - contains deidentified free-text clinical notes collected in the U.S.A. and can act as a good testbed for the latest medical coding models. Searle et al. [155] argued that the defacto gold-standard codes assigned in MIMIC-III had not undergone secondary validation and constituted a silver-standard dataset. The publicly available MIMIC database has promoted research on medical coding. The 2007 Computational Medicine Challenge organized a shared task on ICD coding with a dataset from a hospital in the US [145]. Another recent dataset is CodiEsp in

Spanish, which mainly provides manual codes with in-text explanations (or evidence) of 1,000 Spanish clinical notes, but also English translations of the notes and publications with ICD-10 codes. The dataset was used in the CodiEsp track in eHealth CLEF 2020 [129].

In parallel to the public databases, many studies have also been conducted with private in-house patient notes. For example, Zhang et al. [211] used de-identified medical notes with ICD-10 codes from a hospital in the USA. Teng et al. [165] built a dataset of outpatient medical records with ICD-10 codes collected from a first-class hospital in China. Chen et al. [31] also performed medical coding with datasets with 275,797 EMR documents from two medical departments of top hospitals in China and released a Chinese medical knowledge graph[7]. Rios et al. [150] used a dataset of pathology reports with ICD-0-3 codes collected from the Kentucky Cancer Registry. Liu et al. [117] assigned ICD-10 codes to clinical documents of Dutch and French Datasets. Hansen et al. [69] utilized patients' medication history for diagnosis coding using the Danish register data. Teng et al. [164] introduced several existing datasets for ICD coding. Clinical notes from patient records have strict administrative regulations to protect patient privacy. However, more public de-identified data will help evaluate the generalizability of medical coding models.

4.2 Evaluation Metrics

Medical coding uses evaluation metrics of the multi-label multi-class classification problem to evaluate the predictive performance. Standard evaluation metrics such as the area under the receiver operating characteristic curve (AUC-ROC) and F1-score with two averaging strategies (i.e., micro and macro) and precision at k (P@k) are used by most publications. Micro scores consider all labels jointly and give more weight to frequent labels. The MIMIC-III top-50 dataset contains samples with the top-50 frequent codes. The hierarchical nature of medical codes leads to the need for hierarchical evaluation. Instead of using a flat evaluation that treats each code independently, CoPHE [55] proposes a set of metrics that represent the depth of nodes in a hierarchy, allowing to quantify incorrect but related codes and preserve the counts in the upper layers to assess the issues of under- or over-prediction. Weak Hierarchical Confusion Matrix (WHCM) [56] further adapts the confusion matrix to the document-level multi-label setting to track within-family or out-of-family confusion of codes based on the assumptions of a code "family" in the hierarchy (e.g., diagnosis codes that share the same first three digits of ICD-9).

4.3 Practice in Public Health

Manual medical coding in practice is not perfect; e.g., the overall medium accuracy of coding in the UK was around 83% with a large variance among studies (50-98%) surveyed in [19] around 2012. Errors in manual coding may be due to errors or incompleteness in the patients' data, subjectivity in choosing diagnostic codes, lack of coding expertise, or data entry errors [38]. There are usually backlogs (of months or over a year) of records to be coded [3]. According to the survey in [23], a computer-assisted coding system can potentially help improve coding accuracy, quality, and efficiency; however, the challenges lie in the requirements in the transition from a manual process to a computer-assisted coding environment. Coders should be able to revise the codes suggested by the system and be involved in the system development process [23]. There are also other challenges in linguistics (e.g., hypothetical contexts) and data formats (e.g., hand-written notes). A recent, ongoing project on deploying an NLP system for clinical coding is CogStack for the Artificial Intelligence in Health and Care Award in England [97]. CogStack uses word embedding and concept embedding-based NLP sub-module MedCAT [101] to extract contextual entities with mapping to concepts or codes in UMLS, SNOMED, and ICD-10. We refer readers to [49] for a detailed comment of manual and AI-assisted clinical coding from the perspective of public health in the UK.

[7]https://github.com/PaddlePaddle/Research/tree/master/KG/ACL2020_SignOrSymptom_Relationship

In China, ICD coding is crucial in medical statistics, medical evaluation, medical insurance, billing, and other fields [46]. Currently, the widely used coding rule is ICD-10. The relevant departments regularly update the coding rules and standards and prepare the extended version according to ICD-10, expanding it from 4 bits to 6 bits [176]. However, the vast number of ICD codes and regional differences in the coding system lead to version disparities that adversely impact the coding quality [65]. Moreover, coders must master domain knowledge, coding rules, and medical terminology to complete basic coding tasks [36]. In a Class III hospital in China, only two coders are typically assigned to code ICD for approximately 2000 inpatients daily [84]. Meanwhile, clinical experts have indicated that data loss and distortion are significant issues resulting from imprecise or insufficiently detailed descriptions in clinical notes, posing a major challenge to clinical coding practice. With the continuous development of deep learning technology and hardware, various deep neural network methods for Chinese corpus have been applied to the current ICD automatic coding model [27, 36, 65, 84, 198, 202, 212, 213]. Clinical experts are optimistic about the potential and inevitability of integrating deep learning technology into the clinical practice of ICD coding, which could alleviate the burden on manual coders and improve coding accuracy [35]. However, the deep learning model is challenging to play a decision-making role independently [27]. Clinical coders prefer to believe in the model with solid interpretability and high accuracy [187]. To address this, clinical experts recommend building an ICD coding assistant system that provides queries, recommendations, prompts, and other functionalities, integrating medical knowledge and rules into the clinical coding model based on deep learning [207]. In particular, because the subjective clinical records are not rigorous and complete, it may be advantageous to conduct learning and reasoning based on objective and factual data (such as microbiological events and subscriptions) [118].

There are more case studies in other countries. In Finland, the use of ICD-10 codes is a widely adopted practice, with most healthcare providers generating structured diagnosis codes as part of their day-to-day operations. Medical coding is essential for the purposes of monitoring the quality of clinical care, billing, insurance processing, and clinical research [100]. In the Finnish context, most of the medical coding is carried out within EMR systems, with 100% coverage reached in 2007, and 74% of healthcare operators managing at least 90% of their referral exchange electronically today [149]. The high degree of standardization and digitization of records has made it easy to develop interoperable automated medical coding systems in the Finnish healthcare domain. NLP-based medical coding has been explored as a way to identify unnoticed medical conditions, such as sleeplessness and anxiousness disorders, from clinical notes. The Finnish Institute for Health and Welfare (THL) calls for more widespread structuring of medical information, as well as for a systematic assessment of automated systems for coding and recording purposes, paving way for more widespread use of automated medical coding [76]. In Thailand, Ponthongmak et al. [146] developed deep learning models for ICD-10 coding. In Germany, the statutory health insurance billing for outpatient care is based on the German Uniform Assessment Standard (EBM). Oberste et al. [136] designed an ML system for EBM coding and achieved advanced predictive performance. This review provides a general introduction to real-world applications. However, as a technical review, we could not cover every detailed aspect in the real world. We refer readers to the original publications for details. When applying deep learning-based medical coding methods to country-specific scenarios, the data distribution, languages, and coding schemes changed. Previous performant models that achieved good accuracy on the English MIMIC-III dataset might not work well with new datasets. For example, Ponthongmak et al. [146] showed PLM-ICD and CNN-PubMedBERT performed well in the Thai datasets and translation can enable the adaptation of the pretrained models to Thai-English clinical text. When moving to another country, the results might vary. This review serves as a good guideline for researchers and developers to develop their own deep learning models for medical coding. To comprehensively address the specific nuances of medical coding practices in different countries — for example, identifying the problems solved by country-specific studies, understanding their differences, and examining the empirical results obtained — a dedicated quantitative survey focusing on each country's context would be more suitable.

4.4 Open Sources and Tools

This section introduces useful resources, including open-source software, model implementations, pretrained language models, and medical ontologies.

Open-source codes facilitate reproducible research on clinical NLP algorithms. The Python programming language and the Pytorch deep learning framework dominate the landscape of deep learning-based medical coding. Aitziber et al. [7] developed a machine learning-based extraction system called DTEncoding to automatically generate diagnostic terms (DTs) from electronic health records. SNOMED CT Browser[8] enables users to look up different SNOMED CT editions and clinical healthcare terminologies based on the SNOMED member countries. The National Center for Health Statistics (NCHS) developed and maintained the Mortality Medical Data System (MMDS) to automatically provide classification, entry, and cause-of-death information on death certificates. Clinical-Coder [27] is an online system that assigns ICD-10 codes to Chinese clinical notes. It develops a Dilated Convolutional Attention network with N-gram Matching Mechanism (DACNM) for automated medical coding. Its dilated convolution is the same as that used in DCAN [77]. It further augments the system with an explicit n-gram matching to capture the explicit semantic information. A video demonstration is available [9]. The Clinical Classifications Software (CCS) is a maintained coding system projecting ICD-9-CM codes into coarse-grained CCS codes, which can be used as ontology and disease classification codes. Recently, AnEMIC [95] has been released typically for automatic ICD coding. It is an open-source framework that enables error-reduced data preprocessing, model training, and evaluation for automatic ICD coding. The current release includes multiple convolutional neural networks and transformer-based models.

Much literature [4, 75, 78] shows that models suffer from performance degeneration without considering the in-domain adaptation. Most pretrained language models use text in the majority languages such as English and Spanish for self-supervised pretraining. The research community should also pay more attention to the minority languages and learning algorithms in the low-resource setting.

5 DISCUSSION AND FUTURE DIRECTIONS

Clinical notes generated by clinicians contain rich information about patients' diagnoses and treatment procedures. Healthcare institutions digitized these clinical texts into EHRs and other structural medical and treatment histories of patients for clinical data management, health condition tracking, and automation. This paper reviews deep neural network-based methods for automated medical coding under a unified framework. The advances in deep learning models have significantly improved predictive performance. However, the current trend of leaderboard-oriented research is also concerning. It is easy to fall into the pitfall of excessive neural architecture engineering by chasing the scores on the leaderboard of public benchmarks but missing other critical matters. For example, the widely used MIMIC-III dataset may be imperfectly labeled, or under-coded [155]. Moreover, it only reflects clinical coding practice in the US more than a decade ago (till 2012) and only contains clinical notes written in English. This section summarizes the current research and discusses some critical issues that should be carefully considered when improving predictive performance.

5.1 Summary and Discussion

Our review proposes a unified framework for automated medical coding and categorizes building blocks under the unified framework. First, we introduce four main neural encoder modules, i.e., recurrent neural networks, convolutional neural networks, neural attention mechanism, and graph neural networks. Hierarchical encoders that utilize the hierarchical structure of the text are also discussed in detail. RNNs can capture the sequential dependency in text and are intensively used as the text encoder. Many models also use CNNs to extract local

[8] https://browser.ihtsdotools.org/
[9] Available at https://youtu.be/U4TImTwEysE

features, which is validated to be effective in medical coding with many labels. The neural attention mechanism, especially the self-attention-based Transformer network, suffers from quadratic complexity to the length of the input sequence. A recent study on hierarchical neural architecture with Transformer-based encoders shows that better construction of hierarchical text structure can improve the coding performance [209]. More efficient attention mechanisms such as Longformer [12] and Big Bird [205] for long sequences are also attracting much research attention while still very resource- and time-demanding compared to CNN models.

Most models stack neural layers to build deeper architectures. This review then introduces mechanisms to build deep neural networks. The most frequently adopted method is the residual connection initially proposed for image processing. The residual networks can avoid performance degradation of deep neural architectures and have been used in many CNN-based models. However, the highway networks that use a gating mechanism to control the information flow of deep networks have not been used in building deep models for automated medical coding.

Automated medical coding as a multi-label multi-class classification problem relies on powerful decoder modules to boost predictive performance. The most widely used decoder is the label attention mechanism that learns label-aware representations for medical code prediction. The hierarchical nature of medical classification systems leads to the development of various hierarchical decoders. Other advances, such as multitask learning and few-shot/zero-shot learning-based decoders, are also promising directions.

Finally, we review how to utilize auxiliary information to improve medical coding performance. Auxiliary information includes Wikipedia articles, code descriptions, code hierarchy, chart data, entities, and concepts. Most methods use external text such as medical code descriptions and Wikipedia articles to enhance textual feature learning. Auxiliary information, such as code description, can also act as the regularization for model optimization. For example, the DR-CAML [131] uses the embeddings of ICD code description as a regularizer to enhance the representation learning for rare codes. The code hierarchy connects to hierarchical decoders closely. Chart data and medical imaging data facilitate multimodal learning. Entities and concepts enable knowledge-aware representation learning. External knowledge contributes to robust few-shot and zero-shot medical coding significantly.

5.2 Future Directions

Deep learning has boosted the development of automated medical coding methods. Nevertheless, many challenges exist. This section points out some future directions as follows.

Long-term Dependency and Scalability. It is challenging for neural encoders to capture long-term dependency, especially when clinical notes are extremely long documents. Self-attention-based models that succeed in sentence understanding have scalability issues due to the complexity of self-attention. Although some remedies attempt to make self-attention more efficient in the NLP community, few studies have been done in the context of medical coding. Also, recent deep learning models are becoming increasingly large. Future work should consider the scalability issue when dealing with long clinical documents and high-dimensional medical codes.

Clinical Relatedness. Modern neural models can effectively learn textual features for given input texts. We can usually achieve satisfactory performance with a strong classifier and appropriate training. However, whether the encoding model can capture the clinical relatedness between different text mentions for medical coding is still unclear. Besides, human clinical coders refer to different data types when assigning codes. Multimodal deep learning methods are introduced to learn embeddings of multimodal data. Multimodal alignment and fusion are critical components to capture clinical relatedness across different modalities. Future work needs to deeply infuse clinical knowledge (e.g., knowledge graphs [30, 81]) into the neural encoders, enhancing the model's capability to learn knowledge-aware features and the model's reasoning ability.

Class Imbalance and Hierarchical Decoding. The medical coding tasks suffer from class imbalance with a long tail of rare diagnoses in the class distribution. Nevertheless, current research considers less about the class imbalance issue, which should be addressed in future work. In particular, more effective neural decoders would be required for robust medical coding. The code hierarchy as prior human knowledge sheds light on the imbalanced classes. However, how to enable global and local learning for the whole hierarchy and local branches in the hierarchy is a challenging future work when developing hierarchical decoding approaches. More importantly, enabling few-shot and zero-shot learning for rare and unseen codes without external knowledge is an unsolved problem.

Interpretability. Existing models with a certain level of explanation are post-hoc studies, for example, by interpreting the predictions through the visualization of attention weights [51, 61]. It is important to understand the model's prediction and prioritize features learned by the model. For example, embedding external medical knowledge bases like the unified medical language system might be further utilized to learn rich knowledge-aware representation to help medical text understanding. Knowledge-aware reasoning will need to be introduced to improve the interpretability of medical prediction. However, the medical coding model, mainly occupied by deep learning, is still largely a black box. Thus, further work should focus more on interpretability that can improve the transparency of neural medical coding models, explain and justify model predictions, and ensure accountability and adherence to privacy and ethical guidelines.

Human-in-the-loop Systems. Medical coding models are supposed to facilitate current workflows at hospitals. One important future research direction is to integrate the human-in-the-loop systems [206] into medical coding, in which human experts can interact with the model training process and enhance the model performance [71]. For example, the active learning paradigm can select informative samples for human annotators when preparing the training data. Hospitals usually provide coding guidelines for human coders. For example, one rule from the current guideline requires that hypertension with pregnancy should not be coded as hypertension. Involving human experts can explicitly inject those coding rules into the labeled data and correct the model predictions if they fail to follow the coding guidelines.

Updated Guidelines and Data Shift. Coding guidelines are usually updated frequently. The changes in guidelines should be considered in developing automated medical coding tools to facilitate the updated workflows at hospitals. As time goes by, clinical practice can change. For example, a new pandemic might lead to significant changes in the health systems. Medical coding models should also be able to be robust to data shifts. Considering the increasing nature of health records, incremental learning or lifelong learning [140] might also be studied. Multitask learning [210] that solves the medical coding with different coding schemes can also be further deployed. When new codes enter the standard classification system during the update of coding guidelines, the medical coding model should be adaptable to the zero-shot coding problem. Furthermore, medical coding models should also produce uncertainty-aware predictions when facing updated coding guidelines and schemes.

Novel Encoder-decoder Architectures and Large Language Models. While the existing work can be generalized to an encoder- decoder architecture, sequence-to-sequence (seq2seq) models are less explored for medical coding by modeling text sequence input to code sequence output. Seq2seq models have been applied for multi-label classification to model the correlation among labels [196], and more recently for entity recognition and linking from texts [25]. Atutxa et al. [5] performed a sequence-to-sequence benchmark for ICD-10 coding, published as working notes. However, the paper did not reveal details about how ICD codes are decoded in a sequence-to-sequence manner[10]. Motivated by the seq2seq machine translation model, Atutxa et al. [6] continued the

[10]The decoder generates the output sequence one element at a time. At each step, it takes the previously generated element and the current state as input to predict the next element in the sequence.

study on a real seq2seq model, in which the decoder process considers the previous output when decoding a new medical code. As a future direction, novel decoding processes of the seq2seq method can be explored to better capture the dependency of medical codes. Considering the prominent generation capabilities enabled by large pretrained language models [197] (e.g., encoder-decoder based T5 [148] and decoder-only BioGPT [124]) and instruction-tuned models such as ChatGPT[11], our initial version of this review argued that a potential direction is to generate codes from input documents and prompt-based learning that leverages crafted or learned templates (or prompts) for generation, as surveyed in [116], may also be incorporated to decode medical codes or concepts in the large hierarchical coding space. Recent advances, such as Yang et al. [199], investigated autoregressive generation with prompts. Utilizing the ability of large language models (LLMs), such as ChatGPT, for medical coding would be an interesting direction. Falis et al. [57] studied zero-shot prompting and data augmentation with GPT-3.5 and showed the limited capacity of GPT-3.5. LLM-based methods hold the potentail for contextual understanding on medical text benefiting from the inherent language capabilities of such models. However, they struggle with improving the coding accuracy, especically dealing codes without exisiting examples. Moreover, this direction comes with challenges related to data quality, domain-specific terminology, fine-tuning, and ethical considerations. The integration of LLMs into medical coding represents a challenging path toward improving accuracy and efficiency in this critical healthcare task.

Privacy and Security Concerns. Medical texts often contain both identifying information (such as a patient's name or other identifying characteristics) as well as sensitive information (such as health state or intimate knowledge of their life). Thus, privacy and security concerns must always be addressed when processing, analyzing, and utilizing text-form data. When utilizing machine learning-based models, one must always remember that such models will likely retain patient-specific information unless training data has been thoroughly anonymized. This inherent memorization aspect makes the sharing of models between organizations difficult, as it is arduous to ensure that no patient-specific information can be reverse-engineered from a given model.

6 CONCLUSION

Recent years have witnessed increasing attention to the problem of automated medical coding. This paper reviews automated medical coding from an in-depth perspective that unifies a great variety of existing deep learning-based models into an encoder-decoder framework without losing technical nuances in each specific type of model. Specifically, we discuss 1) neural encoders with recurrent and convolutional networks, neural attention mechanisms, and hierarchical encoders typically used for long clinical notes; 2) mechanisms to build deep architectures, including simple stacking, embedding injection, and residual connection; 3) decoder modules with linear layers, neural attention, hierarchical and multitask decoders; 4) the usage of auxiliary information such as Wikipedia articles, code descriptions, and code hierarchy. Besides, we introduce data for medical coding, the evaluation of medical coding models, and real-world practice. We summarize the limitations and point out future research directions at the end of this review.

ACKNOWLEDGMENTS

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, and grants 336033, 352986, 358246) and EU (H2020 grant 101016775 and NextGenerationEU). H. Dong is supported by Health Data Research UK National Phenomics and Text Analytics Implementation Projects and EPSRC project (EP/V050869/1). The authors acknowledge Matúš Falis for his helpful comments.

[11]https://openai.com/blog/chatgpt

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Arash Afkanpour, Shabir Adeel, Hansenclever Bassani, Arkady Epshteyn, Hongbo Fan, Isaac Jones, Mahan Malihi, Adrian Nauth, Raj Sinha, Sanjana Woonna, Shiva Zamani, Elli Kanal, Mikhail Fomitchev, and Donny Cheung. 2022. BERT for Long Documents: A Case Study of Automated ICD Coding. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Alberto Lavelli, Eben Holderness, Antonio Jimeno Yepes, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 100–107. https://doi.org/10.18653/v1/2022.louhi-1.12

[3] Vera Alonso, João Vasco Santos, Marta Pinto, Joana Ferreira, Isabel Lema, Fernando Lopes, and Alberto Freitas. 2020. Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. *Journal of medical systems* 44, 3 (2020), 1–8.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. ACL, 72–78.

[5] Aitziber Atutxa, Arantza Casillas, Nerea Ezeiza, Víctor Fresno, Iakes Goenaga, Koldo Gojenola, Raquel Martínez, Maite Oronoz Anchordoqui, and Olatz Perez-de Viñaspre. 2018. IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence Approach.. In *CLEF (Working Notes)*. 1.

[6] Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Olatz Perez-de Viñaspre. 2019. Interpretable deep learning to map diagnostic texts to ICD-10 codes. *International journal of medical informatics* 129 (2019), 49–59.

[7] Aitziber Atutxa, Alicia Pérez, and Arantza Casillas. 2017. Machine learning approaches on diagnostic term encoding with the ICD for clinical documentation. *IEEE Journal of Biomedical and Health Informatics* 22, 4 (2017), 1323–1329.

[8] Franz Baader, Ian Horrocks, Carsten Lutz, and Uli Sattler. 2017. *A Basic Description Logic*. Cambridge University Press, Cambridge, 10–49. https://doi.org/10.1017/9781139025355.002

[9] Tian Bai and Slobodan Vucetic. 2019. Improving Medical Code Prediction from Clinical Text via Incorporating Online Knowledge Sources. In *The World Wide Web Conference*. ACM, 72–82.

[10] Weidong Bao, Hongfei Lin, Yijia Zhang, Jian Wang, and Shaowu Zhang. 2021. Medical code prediction via capsule networks and ICD knowledge. *BMC Medical Informatics and Decision Making* 21, 2 (2021), 1–12.

[11] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label Classification of Patient Notes: Case Study on ICD Code Assignment. In *AAAI Workshop*. AAAI, 1–8.

[12] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[13] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3889–3900. https://doi.org/10.18653/v1/2022.acl-long.269

[14] Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. TransICD: Transformer Based Code-Wise Attention Model for Explainable ICD Coding. In *International Conference on Artificial Intelligence in Medicine*. Springer, 469–478.

[15] Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label ICD classification: Sensitivity to hospital service and time. *IEEE Access* 8 (2020), 183534–183545.

[16] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[17] Svetla Boytcheva. 2011. Automatic matching of ICD-10 codes to diagnoses in discharge letters. In *Proceedings of the Second Workshop on Biomedical Natural Language Processing*. ACL, 11–18.

[18] Kristie Burks, Jessie Shields, Joseph Evans, Jodi Plumley, Jarrett Gerlach, and Susan Flesher. 2022. A systematic review of outpatient billing practices. *SAGE Open Medicine* 10 (2022), 20503121221099021.

[19] E.M. Burns, E. Rigby, R. Mamidanna, A. Bottle, P. Aylin, P. Ziprin, and O.D. Faiz. 2011. Systematic review of discharge coding accuracy. *Journal of Public Health* 34, 1 (07 2011), 138–148. https://doi.org/10.1093/pubmed/fdr054

[20] Elaine M Burns, E Rigby, R Mamidanna, A Bottle, P Aylin, P Ziprin, and OD Faiz. 2012. Systematic review of discharge coding accuracy. *Journal of public health* 34, 1 (2012), 138–148.

[21] Erik Cambria, Tim Benson, Chris Eckl, and Amir Hussain. 2012. Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications* 39, 12 (2012), 10533–10543.

[22] Erik Cambria, Amir Hussain, Tariq Durrani, Catherine Havasi, Chris Eckl, and James Munro. 2010. Sentic computing for patient centered applications. In *IEEE 10th International Conference on Signal Processing Proceedings*. IEEE, 1279–1282.

[23] Sharon Campbell and Katrina Giadresco. 2020. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Health Information Management Journal* 49, 1 (2020), 5–18.

[24] Susan E Campbell, Marion K Campbell, Jeremy M Grimshaw, and Anne E Walker. 2001. A systematic review of discharge coding accuracy. *Journal of Public Health* 23, 3 (2001), 205–211.

[25] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*. 1–20. https://openreview.net/forum?id=5k8F6UU39V

[26] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 3105–3114.

[27] Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL, 294–301.

[28] Rich Caruana. 1997. Multitask learning. *Machine Learning* 28 (1997), 41–75.

[29] Finneas Catling, Georgios P Spithourakis, and Sebastian Riedel. 2018. Towards automated clinical coding. *International Journal of Medical Informatics* 120 (2018), 50–61.

[30] S. Chari, D. M. Gruen, O. Seneviratne, and D. L. McGuinness. 2020. Directions for Explainable Knowledge-Enabled Systems. In *Knowledge Graphs for eXplainable AI – Foundations, Applications and Challenges*, Ilaria Tiddi, Freddy Lecue, and Pascal Hitzler (Eds.). Vol. 47. IOS Press, Amsterdam, 245.

[31] Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. 2020. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 3143–3153.

[32] Jiaoyan Chen, Yuan He, Yuxia Geng, Ernesto Jiménez-Ruiz, Hang Dong, and Ian Horrocks. 2023. Contextual semantic embeddings for ontology subsumption prediction. *World Wide Web* (2023), 1–23.

[33] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. 2021. OWL2Vec*: Embedding of OWL ontologies. *Machine Learning* 110, 7 (2021), 1813–1845.

[34] Pei-Fu Chen, Tai-Liang He, Sheng-Che Lin, Yuan-Chia Chu, Chen-Tsung Kuo, Feipei Lai, Ssu-Ming Wang, Wan-Xuan Zhu, Kuan-Chih Chen, Lu-Cheng Kuo, et al. 2022. Training a Deep Contextualized Language Model for International Classification of Diseases, 10th Revision Classification via Federated Learning: Model Development and Validation Study. *JMIR Medical Informatics* 10, 11 (2022), e41342.

[35] Pei-Fu Chen, Ssu-Ming Wang, Wei-Chih Liao, Lu-Cheng Kuo, Kuan-Chih Chen, Yu-Cheng Lin, Chi-Yu Yang, Chi-Hao Chiu, Shu-Chih Chang, Feipei Lai, et al. 2021. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. *JMIR Medical Informatics* 9, 8 (2021), e23230.

[36] YunZhi Chen, HuiJuan Lu, and LanJuan Li. 2017. Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PloS One* 12, 3 (2017), e0173410.

[37] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. PMLR, 301–318.

[38] E. Coiera. 2015. Chapter 24: Natural language and formal terminology. In *Guide to Health Informatics*. CRC Press.

[39] E. Coiera. 2015. *Guide to Health Informatics*. CRC Press, Taylor & Francis Group, Boca Raton, Chapter Chapter 23 Healthcare terminologies and classification systems, 381–399. https://doi.org/10.1201/b13617

[40] Isabel Coutinho and Bruno Martins. 2022. Transformer-based models for ICD-10 coding of death certificates with Portuguese text. *Journal of Biomedical Informatics* 136 (2022), 104232.

[41] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Talukdar, and Steven Carroll. 2007. Automatic Code Assignment to Medical Text. In *Proceedings of BioNLP: Biological, Translational, and Clinical language processing*. ACL, 129–136.

[42] Christine M Cutillo, Karlie R Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D Mandl, and MI in Healthcare Workshop Working Group Beck Tyler 1 Collier Elaine 1 Colvis Christine 1 Gersing Kenneth 1 Gordon Valery 1 Jensen Roxanne 8 Shabestari Behrouz 9 Southall Noel 1. 2020. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine* 3, 1 (2020), 47.

[43] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting Transformer-based Models for Long Document Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7212–7230. https://aclanthology.org/2022.findings-emnlp.534

[44] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2978–2988.

[45] Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the International Conference on Information and Knowledge Management*. 132–139.

[46] Xiaolin Diao, Yanni Huo, Shuai Zhao, Jing Yuan, Meng Cui, Yuxin Wang, Xiaodan Lian, and Wei Zhao. 2021. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *International Journal of Medical Informatics* 153 (2021), 104543.

[47] Molla S Donaldson et al. 1999. Measuring the quality of health care. (1999).
[48] Hang Dong, Jiaoyan Chen, Yuan He, and Ian Horrocks. 2023. Ontology Enrichment from Texts: A Biomedical Dataset for Concept Discovery and Placement. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 5316–5320. https://doi.org/10.1145/3583780.3615126
[49] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matteson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated Clinical Coding: What, Why, and Where We Are? *npj Digital Medicine* 5 (2022), 1–8. Issue 159.
[50] Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable Automated Coding of Clinical Notes using Hierarchical Label-wise Attention Networks and Label Embedding Initialisation. *Journal of Biomedical Informatics* 116 (2021), 103728.
[51] Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of Biomedical Informatics* 116 (2021), 103728.
[52] Hang Dong, Víctor Suárez-Paniagua, Huayu Zhang, Minhong Wang, Arlene Casey, Emma Davidson, Jiaoyan Chen, Beatrice Alex, William Whiteley, and Honghan Wu. 2023. Ontology-Based and Weakly Supervised Rare Disease Phenotyping from Clinical Notes. *BMC Medical Informatics and Decision Making* 86 (2023). Issue 23.
[53] Junwen Duan, Han Jiang, and Ying Yu. 2023. MHLAT: Multi-Hop Label-Wise Attention Model for Automatic ICD Coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[54] Duodecim. 2023. Current Care Guidelines. https://www.kaypahoito.fi/
[55] Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2021. CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 907–912.
[56] Matúš Falis, Hang Dong, Alexandra Birch, and Beatrice Alex. 2022. Horses to Zebras: Ontology-Guided Data Augmentation and Synthesis for ICD-9 Coding. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Dublin, Ireland, 389–401.
[57] Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can GPT-3.5 Generate and Code Discharge Summaries? arXiv:2401.13512 [cs.CL]
[58] Matus Falis, Maciej Pajak, Aneta Lisowska, Patrick Schrempf, Lucas Deckers, Shadia Mikhael, Sotirios Tsaftaris, and Alison O'Neil. 2019. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. ACL, 168–177.
[59] Richárd Farkas and György Szarvas. 2008. Automatic Construction of Rule-based ICD-9-CM Coding Systems. In *BMC Bioinformatics*, Vol. 9(Suppl 3). Springer, 1–9.
[60] Martha Dais Ferreira, Michal Malyska, Nicola Sahar, Riccardo Miotto, Fernando Paulovich, and Evangelos Milios. 2021. Active learning for medical code assignment. In *ACM Conference on Health, Inference, and Learning (CHIL) Workshop*.
[61] Malte Feucht, Zhiliang Wu, Sophia Althammer, and Volker Tresp. 2021. Description-based Label Attention Classifier for Explainable ICD-9 Classification. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. ACL, 62–66.
[62] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.
[63] Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia Tourassi. 2021. Limitations of Transformers on Clinical Text Classification. *IEEE Journal of Biomedical and Health Informatics* 25, 9 (2021), 3596–3607. https://doi.org/10.1109/JBHI.2021.3062322
[64] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. (Aug. 2021), 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295
[65] Yue Gao, Xiangling Fu, Xien Liu, and Ji Wu. 2021. Multi-features-Based Automatic Clinical Coding for Chinese ICD-9-CM-3. In *Proceedings of the 30th International Conference on Artificial Neural Networks and Machine Learning*. Springer, 473–486.
[66] Xueren Ge, Ronald Dean Williams, John A Stankovic, and Homa Alemzadeh. 2023. DKEC: Domain Knowledge Enhanced Multi-Label Classification for Electronic Health Records. *arXiv preprint arXiv:2310.07059* (2023).
[67] Gonçalo Gomes, Isabel Coutinho, and Bruno Martins. 2024. Accurate and Well-Calibrated ICD Code Assignment Through Attention Over Diverse Label Embeddings. In *Procceedings of EACL*.
[68] Irit Hadar and Pnina Soffer. 2006. Variations in conceptual modeling: classification and ontological analysis. *Journal of the Association for Information Systems* 7, 8 (2006), 1.
[69] Emil Riis Hansen, Tomer Sagi, Katja Hose, Gregory YH Lip, Torben Bjerregaard Larsen, and Flemming Skjøth. 2022. Assigning diagnosis codes using medication history. *Artificial Intelligence in Medicine* 128 (2022), 102307.
[70] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.
[71] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[72] Wen-hui Hou, Xiao-kang Wang, Ya-nan Wang, Jian-qiang Wang, and Fei Xiao. 2024. Modelling long medical documents and code associations for explainable automatic ICD coding. *Expert Systems with Applications* (2024), 123519.
[73] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7132–7141.
[74] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. 2022. Transformer quality in linear time. In *International Conference on Machine Learning*. PMLR, 9099–9117.
[75] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).
[76] Hannele Hyppönen, Riikka Vuokko, Persephone Doupi, and Päivi Mäkelä-Bengs. 2014. Sähköisen potilaskertomuksen rakenteistaminen: Menetelmät, arviointikäytännöt ja vaikutukset. (2014).
[77] Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated Convolutional Attention Network for Medical Code Assignment from Clinical Text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. ACL, 73–78.
[78] Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the Magic of BERT Apply to Medical Code Assignment? A Quantitative Study. *Computers in Biology and Medicine* 139 (2021), 104998.
[79] Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal Ideation and Mental Disorder Detection with Attentive Relation Networks. *Neural Computing and Applications* 34 (2022), 10309–10319. Issue 13.
[80] Shaoxiong Ji and Pekka Marttinen. 2023. Patient Outcome and Zero-shot Diagnosis Prediction with Hypernetwork-guided Multitask Learning. In *Proceedings of EACL*.
[81] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2022. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *IEEE Transactions on Neural Networks and Learning Systems* 33 (2022), 494–514. Issue 2.
[82] Shaoxiong Ji, Shirui Pan, and Pekka Marttinen. 2021. Medical Code Assignment with Gated Convolution and Note-Code Interaction. In *Findings of ACL-IJCNLP*. ACL, 1034–1043.
[83] Menglin Jia, Austin Reiter, Ser-Nam Lim, Yoav Artzi, and Claire Cardie. 2021. When in Doubt: Improving Classification Performance with Alternating Normalization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. ACL, 1716–1723.
[84] Zheng Jia, Weifeng Qin, Huilong Duan, Xudong Lv, and Haomin Li. 2017. A hybrid method for ICD-10 auto-coding of Chinese diagnoses. In *MEDINFO 2017: Precision Healthcare through Informatics*. IOS Press, 427–431.
[85] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data* 3 (2016), 160035.
[86] Pollard Tom Horng Steven Celi Leo Anthony Johnson, Alistair and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet. https://doi.org/10.13026/1n74-ne17
[87] Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In *Proceedings of BioNLP*. ACL, 328–332.
[88] Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2021. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *arXiv preprint arXiv:2107.10652* (2021).
[89] Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2023. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Systems with Applications* 213 (2023), 118997.
[90] Aparup Khatua, Apalak Khatua, and Erik Cambria. 2019. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management* 56, 1 (2019), 247–257.
[91] Sarika R Khope and Susan Elias. 2023. Strategies of Predictive Schemes and Clinical Diagnosis for Prognosis Using MIMIC-III: A Systematic Review. In *Healthcare*, Vol. 11. Multidisciplinary Digital Publishing Institute, 710.
[92] Byung-Hak Kim, Zhongfen Deng, Philip S Yu, and Varun Ganapathi. 2022. Can Current Explainability Help Provide References in Clinical Notes to Support Humans Annotate Medical Codes?. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis*.
[93] Byung-Hak Kim and Varun Ganapathi. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Machine Learning for Healthcare Conference*. PMLR, 196–208.
[94] Daeseong Kim, Haanju Yoo, and Sewon Kim. 2022. An Automatic ICD Coding Network Using Partition-Based Label Attention. *arXiv preprint arXiv:2211.08429* (2022).
[95] Juyong Kim, Abheesht Sharma, Suhas Shanbhogue, Pradeep Ravikumar, and Jeremy C Weiss. 2022. AnEMIC: A Framework for Benchmarking ICD Coding Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), System Demonstrations*. ACL.
[96] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[97] King's College Hospital. 2021. CogStack wins an Artificial Intelligence in Health and Care. https://www.kch.nhs.uk/news/public/news/view/34965.
[98] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.
[99] Hannu Kivijärvi and Karoliina Pärnänen. 2023. Instrumental usability and effective user experience: Interwoven drivers and outcomes of Human-Computer interaction. *International Journal of Human-Computer Interaction* 39, 1 (2023), 34–51.
[100] Jorma Komulainen. 2012. Suomalainen tautien kirjaamisen ohjekirja. (2012).
[101] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif. Intell. Med.* 117 (July 2021), 102083. https://doi.org/10.1016/j.artmed.2021.102083
[102] Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, and Robert Hoehndorf. 2019. EL Embeddings: Geometric construction of models for the description logic EL++. In *International Joint Conferences on Artificial Intelligence*. 6103–6109.
[103] Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 2627–2636. https://doi.org/10.18653/v1/2021.naacl-main.208
[104] Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI, 8180–8187.
[105] Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, et al. 2022. Neural natural language processing for unstructured data in electronic health records: a review. *Computer Science Review* 46 (2022), 100511.
[106] Xingwang Li, Yijia Zhang, Deshi Dong, Hao Wei, Mingyu Lu, et al. 2021. JLAN: medical code prediction via joint learning attention networks and denoising mechanism. *BMC Bioinformatics* 22, 1 (2021), 1–21.
[107] Xiaobo Li, Yijia Zhang, Xingwang Li, Xianwei Pan, Jian Wang, and Mingyu Lu. 2023. Automatic International Classification of Diseases Coding via Note-Code Interaction Network with Denoising Mechanism. *Journal of Computational Biology* 30, 8 (2023), 912–925.
[108] Xiaobo Li, Yijia Zhang, Xingwang Li, Jian Wang, and Mingyu Lu. 2023. NIDN: Medical Code Assignment via Note-Code Interaction Denoising Network. In *Proceedings of 18th International Symposium on Bioinformatics Research and Applications (ISBRA)*. 62–74.
[109] Xinhang Li, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. Towards Automatic ICD Coding via Knowledge Enhanced Multi-Task Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1238–1248.
[110] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*.
[111] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics* 2, 01 (1993), 41–51.
[112] Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. 2008. Large Scale Diagnostic Code Classification for Medical Patient Records. In *Proceedings of IJCNLP*. ACL.
[113] Jie-Jyun Liu, Tsung-Han Yang, Si-An Chen, and Chih-Jen Lin. 2021. Parameter Selection: Why We Should Pay More Attention to It. In *ACL-IJCNLP*. ACL, 825–830.
[114] Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Automated ICD Coding using Extreme Multi-label Long Text Transformer-based Models. *Journal of Biomedical Informatics* 133 (2022), 104161.
[115] Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics* 133 (2022), 104161.
[116] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
[117] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. ACL, 5941–5953.
[118] Zichen Liu, Xuyuan Liu, Yanlong Wen, Guoqing Zhao, Fen Xia, and Xiaojie Yuan. 2022. TreeMAN: Tree-enhanced Multimodal Attention Network for ICD Coding. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3054–3063.
[119] Chang Lu, Chandan Reddy, Ping Wang, and Yue Ning. 2024. Towards Semi-Structured Automatic ICD Coding via Tree-based Contrastive Learning. *Advances in Neural Information Processing Systems* 36 (2024).
[120] Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. 2020. Multi-label Few/Zero-shot Learning with Knowledge Aggregated from Multiple Label Graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2935–2943.

[121] Pengli Lu and Jingjin Xue. 2023. Combining transformer-based model and GCN to predict ICD codes from clinical records. *Knowledge-Based Systems* 282 (2023), 111113.
[122] Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. CoRelation: Boosting Automatic ICD Coding Through Contextualized Code Relation Learning. In *Proceedings of LREC-COLING*.
[123] Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: Towards Automated ICD Coding via Feature Compression. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. ACL, 2096–2101.
[124] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (2022).
[125] Julia Medori and Cédrick Fairon. 2010. Machine Learning and Features Selection for Semi-automatic ICD-9-CM Encoding. In *Proceedings of LOUHI Workshop on Text and Data Mining of Health Documents*. ACL, 84–89.
[126] Genevieve B Melton and George Hripcsak. 2005. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association* 12, 4 (2005), 448–457.
[127] George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. 2022. ICDBigBird: A Contextual Embedding Model for ICD Code Classification. In *Proceedings of Biomedical Natural Language Processing*. 330–336.
[128] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*. 3111–3119.
[129] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.
[130] Mark Morsch. 2010. Computer-assisted coding: the secret weapon. CAC does not eliminate the need for medical-coding professionals to be involved in the coding process, but it can make them more productive and accurate. *Health Management Technology* 31, 2 (2010), 24–26.
[131] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of NAACL-HLT*. 1101–1111.
[132] G Jaya Nair. 2013. Ensuring quality in the coding process: A key differentiator for the accurate interpretation of safety data. *Perspectives in clinical research* 4, 3 (2013), 181.
[133] Boon Liang Clarence Ng, Diogo Santos, and Marek Rei. 2023. Modelling Temporal Document Sequences for Clinical ICD Coding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 1640–1649. https://doi.org/10.18653/v1/2023.eacl-main.120
[134] Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, and Stefan Winkler. 2023. A Two-Stage Decoder for Efficient ICD Coding. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4658–4665. https://doi.org/10.18653/v1/2023.findings-acl.285
[135] Kunying Niu, Yifan Wu, Yaohang Li, and Min Li. 2023. Retrieve and rerank for automated ICD coding via Contrastive Learning. *Journal of Biomedical Informatics* 143 (2023), 104396.
[136] Luis Oberste, Nikola Finze, Philipp Hoffmann, and Armin Heinzl. 2022. Supporting the Billing Process in Outpatient Medical Care: Automated Medical Coding Through Machine Learning. In *Enropean Conference on Information Systems*. 1–18.
[137] Luis Oberste and Armin Heinzl. 2022. User-centric explainability in healthcare: a knowledge-level perspective of informed machine learning. *IEEE Transactions on Artificial Intelligence* (2022).
[138] International Health Terminology Standards Development Organisation. 2024. Clinical Finding Defining Attributes. SNOMED CT Editorial Guide https://confluence.ihtsdotools.org/display/DOCEG/Clinical+Finding+Defining+Attributes. Accessed: March 2024.
[139] Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624.
[140] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks* 113 (2019), 54–71.
[141] Jong-Ku Park, Ki-Soon Kim, Tae-Yong Lee, Kang-Sook Lee, Duk-Hee Lee, Sun-Hee Lee, Sun-Ha Jee, Il Suh, Kwang-Wook Koh, So-Yeon Ryu, et al. 2000. The accuracy of ICD codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine and Public Health* 33, 1 (2000), 76–82.
[142] Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based Automatic ICD Coding: Limitations and Opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. 54–63.
[143] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*. 1532–1543.
[144] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *JAMIA* 21, 2 (2014), 231–237.

[145] John Pestian, Chris Brew, Pawel Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*. 97–104.
[146] Wanchana Ponthongmak, Ratchainant Thammasudjarit, Gareth J McKay, John Attia, Nawanan Theera-Ampornpunt, and Ammarin Thakkinstian. 2023. Development and external validation of automated ICD-10 coding from discharge summaries using deep learning approaches. *Informatics in Medicine Unlocked* (2023), 101227.
[147] Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed Memory Networks for Clinical Diagnostic Inferencing. In *Proceedings of AAAI*.
[148] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html
[149] Jarmo Reponen, Niina Keränen, Ronja Ruotanen, Timo Tuovinen, Jari Haverinen, and Maarit Kangas. 2021. Tieto-ja viestintäteknologian käyttö terveydenhuollossa vuonna 2020: Tilanne ja kehityksen suunta. (2021).
[150] Anthony Rios, Eric B Durbin, Isaac Hands, and Ramakanth Kavuluru. 2021. Assigning ICD-O-3 codes to pathology reports using neural multi-task training with hierarchical regularization. In *Proceedings of ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 1–10.
[151] Anthony Rios and Ramakanth Kavuluru. 2018. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of EMNLP*. 3132–3142.
[152] Kevin Roitero, Beatrice Portelli, Mihai Horia Popescu, and Vincenzo Della Mea. 2021. DiLBERT: Cheap embeddings for disease related medical NLP. *IEEE Access* 9 (2021), 159714–159723.
[153] Najmeh Sadoughi, Greg P Finley, James Fone, Vignesh Murali, Maxim Korenevski, Slava Baryshnikov, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. Medical code prediction with multi-view convolution and description-regularized label-dependent attention. *arXiv preprint arXiv:1811.01468* (2018).
[154] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A Public-access Intensive Care Unit Database. *Critical Care Medicine* 39, 5 (2011), 952.
[155] Thomas Searle, Zina Ibrahim, and Richard Dobson. 2020. Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset. In *Proceedings of SIGBioMed Workshop on Biomedical Language Processing*. 76–85.
[156] Thomas Searle, Zeljko Kraljevic, Rebecca Bendayan, Daniel Bean, and Richard Dobson. 2019. MedCATTrainer: A Biomedical Free Text Annotation Interface with Active Learning and Research Use Case Specific Customisation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. 139–144.
[157] Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards Automated ICD Coding Using Deep Learning. *arXiv preprint arXiv:1711.04075* (2017).
[158] Gail I Smith and June Bronnert. 2010. Transitioning to CAC: the skills and tools required to work with computer-assisted coding. *Journal of AHIMA* 81, 7 (2010), 60–61.
[159] Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric P Xing. 2020. Generalized Zero-Shot Text Classification for ICD Coding. In *IJCAI*. 4018–4024.
[160] Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A Systematic Literature Review of Automated Clinical Coding and Classification systems. *JAMIA* 17, 6 (2010), 646–651.
[161] Wei Sun, Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2021. Multitask Recalibrated Aggregation Network for Medical Code Prediction. In *Proceedings of ECML-PKDD*. 367–383.
[162] Wei Sun, Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2023. Multitask Balanced and Recalibrated Network for Medical Code Prediction. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2023), 1–20.
[163] Hanna Suominen, Filip Ginter, Sampo Pyysalo, Antti Airola, Tapio Pahikkala, S Salanter, and Tapio Salakoski. 2008. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In *Proceedings of the ICML/UAI/COLT workshop on machine learning for health-care applications*.
[164] Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. 2022. A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[165] Fei Teng, Zheng Ma, Jie Chen, Ming Xiao, and Lufei Huang. 2020. Automatic medical code assignment via deep learning approach for intelligent healthcare. *IEEE JBHI* 24, 9 (2020), 2506–2515.
[166] Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. 2020. Explainable prediction of medical codes with knowledge graphs. *Frontiers in bioengineering and biotechnology* 8 (2020), 867.
[167] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200* (2020).

[168] Shang-Chi Tsai, Ting-Yun Chang, and Yun-Nung Chen. 2019. Leveraging hierarchical category knowledge for data-imbalanced multi-label diagnostic text understanding. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. 39–43.

[169] Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling Diagnostic Label Correlation for Automatic ICD Coding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4043–4052.

[170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[171] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[172] Kaushik P. Venkatesh, Marium M. Raza, and Joseph C. Kvedar. 2023. Automating the overburdened clinical coding system: challenges and next steps. *npj Digital Medicine* 6, 1 (2023), 16. https://doi.org/10.1038/s41746-023-00768-0

[173] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for ICD coding from clinical text. In *Proceedings of IJCAI*.

[174] W3C Recommendation. 2012. OWL EL, OWL 2 Web Ontology Language Profiles (Second Edition). https://www.w3.org/TR/owl2-profiles/#OWL_2_EL. Accessed: 2024-03-13.

[175] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint Embedding of Words and Labels for Text Classification. In *Proceedings of ACL*. 2321–2331.

[176] Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, et al. 2020. A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes. *Journal of Biomedical Informatics* 105 (2020), 103418.

[177] Ran Wang, Siyu Long, Xinyu Dai, Shujian Huang, Jiajun Chen, et al. 2021. Meta-LMTC: Meta-Learning for Large-Scale Multi-Label Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 8633–8646.

[178] Sen Wang, Xiaojun Chang, Xue Li, Guodong Long, Lina Yao, and Quan Z Sheng. 2016. Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE TKDE* 28, 12 (2016), 3191–3202.

[179] Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Huasheng Liang, Qiang Yan, Evangelos Kanoulas, and Maarten de Rijke. 2021. Few-Shot Electronic Health Record Coding through Graph Contrastive Learning. *arXiv preprint arXiv:2106.15467* (2021).

[180] Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jian-Yun Nie, Jun Ma, and Maarten de Rijke. 2020. Coding Electronic Health Records with Adversarial Reinforcement Path Generation. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*. 801–810.

[181] Shi Wang, Daniel Tang, Luchen Zhang, Huilin Li, and Ding Han. 2022. HieNet: Bidirectional Hierarchy Framework for Automated ICD Coding. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part II*. 523–539.

[182] Ssu-ming Wang, Yu-hsuan Chang, Lu-cheng Kuo, Feipei Lai, Yun-nung Chen, Fei-yun Yu, Chih-wei Chen, Zong-wei Li, and Yufang Chung. 2020. Using deep learning for automatic ICD-10 classification from free-text data. *European Journal of Biomedical Informatics* 16, 1 (2020).

[183] Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A Novel Framework Based on Medical Concept Driven Attention for Explainable Medical Code Prediction via External Knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*. 1407–1416.

[184] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *Comput. Surveys* 53, 3 (2020), 1–34.

[185] Sarah Wiegreffe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. 2019. Clinical Concept Extraction for Document-Level Coding. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 261–272.

[186] Wong-Lin, McClean, McCombe, Kaur, Sanchez-Bornot, Gillespie, Todd, Finn, Joshi, Kane, and McGuinness. 2020. Shaping a data-driven era in dementia care pathway through computational neurology approaches. *BMC Medicine* 18, 1 (16 Dec 2020), 398. https://doi.org/10.1186/s12916-020-01841-1

[187] Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. 2022. Model Distillation for Faithful Explanations of Medical Code Predictions. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. 412–425.

[188] Haoran Wu, Wei Chen, Shuang Xu, and Bo Xu. 2021. Counterfactual Supporting Facts Extraction for Explainable Medical Record Based Diagnosis with Graph Network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1942–1955.

[189] Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard JB Dobson. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association* 25, 5 (01 2018), 530–537. https://doi.org/10.1093/jamia/ocx160

[190] Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al. 2023. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning* 16, 2 (2023), 119–328.
[191] Yuzhou Wu, Zhigang Chen, Xin Yao, Xuechen Chen, Zeren Zhou, and Jinkai Xue. 2022. JAN: Joint Attention Networks for Automatic ICD Coding. *IEEE Journal of Biomedical and Health Informatics* 26, 10 (2022), 5235–5246.
[192] Jing Xie, Xin Li, Ye Yuan, Yi Guan, Jingchi Jiang, Xitong Guo, and Xin Peng. 2024. Knowledge-based dynamic prompt learning for multi-label disease diagnosis. *Knowledge-Based Systems* 286 (2024), 111395.
[193] Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of ACM CIKM*. 649–658.
[194] Bo Xiong, Nico Potyka, Trung-Kien Tran, Mojtaba Nayyeri, and Steffen Staab. 2022. Faithful embeddings for EL++ knowledge bases. In *International Semantic Web Conference*. Springer, 22–38.
[195] Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K Khanna, Jacek B Cywinski, Kamal Maheshwari, et al. 2019. Multimodal machine learning for automated ICD coding. In *Machine learning for healthcare conference*. PMLR, 197–215.
[196] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence Generation Model for Multi-label Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3915–3926.
[197] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *npj Digital Medicine* 5, 1 (2022), 194.
[198] Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. 2018. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific reports* 8, 1 (2018), 6329.
[199] Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2023. Multi-label Few-shot ICD Coding as Autoregressive Generation with Prompt. In *AAAI*.
[200] Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge Injected Prompt Based Fine-tuning for Multi-label Few-shot ICD Coding. In *Findings of EMNLP*.
[201] Vithya Yogarajan, Bernhard Pfahringer, Tony Smith, and Jacob Montiel. 2021. Improving Predictions of Tail-end Labels using Concatenated BioMed-Transformers for Long Medical Documents. *arXiv preprint arXiv:2112.01718* (2021).
[202] Ying Yu, Min Li, Liangliang Liu, Zhihui Fei, Fang-Xiang Wu, and Jianxin Wang. 2019. Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN. *Journal of Biomedical Informatics* 91 (2019), 103114.
[203] Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. 2020. The Graph-based Mutual Attentive Network for Automatic Diagnosis. In *IJCAI*. 3393–3399.
[204] Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. In *ACL*.
[205] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Advances in neural information processing systems*, Vol. 33. 17283–17297.
[206] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.
[207] Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. 2019. Automatic ICD-9 Coding via Deep Transfer Learning. *Neurocomputing* 324 (2019), 43–50.
[208] Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing* 18, 6 (1989), 1245–1262.
[209] Ning Zhang and Maciej Jankowski. 2022. Hierarchical BERT for Medical Document Understanding. *arXiv preprint arXiv:2204.09600* (2022).
[210] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE TKDE* 34 (2021), 5586 – 5609. Issue 12.
[211] Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 24–34.
[212] Shuai Zhao, Xiaolin Diao, Yun Xia, Yanni Huo, Meng Cui, Yuxin Wang, Jing Yuan, and Wei Zhao. 2023. Automated ICD coding for coronary heart diseases by a deep learning method. *Heliyon* (2023), e14037.
[213] Lingling Zhou, Cheng Cheng, Dong Ou, and Hao Huang. 2020. Construction of a semi-automatic ICD-10 coding system. *BMC medical informatics and decision making* 20 (2020), 1–12.
[214] Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic ICD Coding via Interactive Shared Representation Networks with Self-distillation Mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5948–5957.

Received 10 June 2022; revised 18 March 2024; accepted 2 May 2024