

## Transformers and large language models in healthcare: A review



Subhash Nerella <sup>a,1</sup>, Sabyasachi Bandyopadhyay <sup>b,1</sup>, Jiaqing Zhang <sup>c</sup>, Miguel Contreras <sup>a</sup>, Scott Siegel <sup>a</sup>, Aysegul Bumin <sup>d</sup>, Brandon Silva <sup>d</sup>, Jessica Sena <sup>e</sup>, Benjamin Shickel <sup>f</sup>, Azra Bihorac <sup>f</sup>, Kia Khezeli <sup>a,\*</sup>, Parisa Rashidi <sup>a</sup>

<sup>a</sup> Department of Biomedical Engineering, University of Florida, Gainesville, United States

<sup>b</sup> Department of Medicine, Stanford University, Stanford, United States

<sup>c</sup> Department of Electrical and Computer Engineering, University of Florida, Gainesville, United States

<sup>d</sup> Department of Computer and Information Science and Engineering, University of Florida, Gainesville, United States

<sup>e</sup> Department Of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

<sup>f</sup> Department of Medicine, University of Florida, Gainesville, United States

### ARTICLE INFO

**Keywords:**  
 Transformers  
 Healthcare  
 Electronic Health Records  
 Large Language Models  
 Medical Imaging  
 Natural Language Processing

### ABSTRACT

With Artificial Intelligence (AI) increasingly permeating various aspects of society, including healthcare, the adoption of the Transformers neural network architecture is rapidly changing many applications. Transformer is a type of deep learning architecture initially developed to solve general-purpose Natural Language Processing (NLP) tasks and has subsequently been adapted in many fields, including healthcare. In this survey paper, we provide an overview of how this architecture has been adopted to analyze various forms of healthcare data, including clinical NLP, medical imaging, structured Electronic Health Records (EHR), social media, biophysiological signals, biomolecular sequences. Furthermore, which have also include the articles that used the transformer architecture for generating surgical instructions and predicting adverse outcomes after surgeries under the umbrella of critical care. Under diverse settings, these models have been used for clinical diagnosis, report generation, data reconstruction, and drug/protein synthesis. Finally, we also discuss the benefits and limitations of using transformers in healthcare and examine issues such as computational cost, model interpretability, fairness, alignment with human values, ethical implications, and environmental impact.

### 1. Introduction

The last decade has seen an explosion in data generated by healthcare practices. Currently, healthcare data accounts for 30% of the global data ecosystem and is expected to grow in the coming years [1]. The increasing availability of digital patient data has enabled the development of machine learning algorithms to support diagnosis, prognosis, and clinical decision-making.

Transformer [2] is a type of Deep Neural Network (DNN) introduced in 2017 for sequence modeling problems, especially in the Natural Language Processing (NLP) domain [3]. Before the introduction of the Transformer [2], the most popular sequential deep learning architectures, such as recurrent neural networks (RNNs) [4] and their variants, worked in a serial fashion which precluded parallelization during training, therefore substantially increasing the training time. In contrast,

Transformer employs parallelizable scaled dot-product attention mechanism. This unique attention mechanism allows for large-scale pretraining. Additionally, self-supervised pretraining on large unlabeled datasets using approaches such as input masking enabled transformers to be trained without costly annotations.

Although originally designed for the NLP [3] domain, Transformers have witnessed adaptations in various domains such as computer vision [5,6], remote sensing [7], time series [8], speech processing [9] and multimodal learning [10]. Consequently, modality-specific surveys emerged, focusing on medical imaging [11–13] and biomedical language models [14]. However, a comprehensive review of all healthcare-oriented literature that has employed the Transformer architecture has not been undertaken. This paper aims to provide a comprehensive review of Transformer models utilized across multiple healthcare data modalities while focusing on notable architectural changes undergone

\* Corresponding author.

E-mail address: [parisa.rashidi@ufl.edu](mailto:parisa.rashidi@ufl.edu) (P. Rashidi).

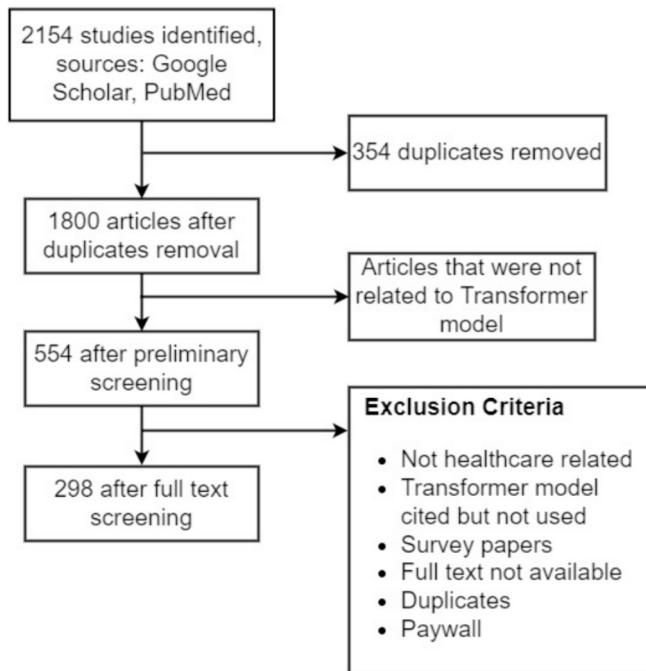
<sup>1</sup> Contributed equally.

by the original Transformer model through this process of evolution. This is critical and timely because the transformer architecture is being rapidly incorporated into almost every healthcare domain; it is critical to understand common patterns and features in this adoption. Furthermore, we discuss pretraining strategies designed to manage the lack of robust and/or annotated healthcare datasets. The rest of the paper is organized as follows: **Section 2** discusses our search strategy; **Section 3** describes the architecture of the original transformer; **Section 4** describes the two most commonly used Transformer variants: the Bidirectional Encoder Representations from Transformers (BERT) and the Vision Transformer (ViT). **Section 5** describes advancements in large language models (LLM), and **Sections 6 through 12** provide a review of Transformers in healthcare. Finally, **Section 13** discusses limitations, interpretability, environmental impact, computational costs, bias, and fairness. This review summarizes the use of transformer-based deep learning models in healthcare and provides a critical analysis of the inherent deficiencies of these models and discusses possible future directions for this field.

## 2. Search strategy and selection criteria

We used Google Scholar and PubMed search engines to search for studies. Since Vaswani et al.'s initial Transformer network [2] was published in 2017, we limited our search to studies published after 2017. We also limited our search to studies published before March 2023 to complete this review. The extraction process and exclusion criteria is shown in Fig. 1. The search was divided into six categories: clinical NLP, EHR, social media, medical imaging, biomolecules, and bio-physical signals.

For each category, we used the terms "health" or "medical" or "clinical" to focus the search on the healthcare domain. Finally, each category used a precise set of keywords unique to that domain. The keywords were combined with logical operators such as "AND" and "OR" to enhance search fidelity. A detailed list of search queries can be found in Table 1. We used Harzing's Publish or Perish [15] to retrieve studies and Covidence [16] to select relevant studies. For most medical domains an article that was preliminarily short-listed was reviewed by



**Fig. 1.** Flow diagram depicting the process for selecting relevant studies for inclusion and exclusion.

**Table 1**  
Search queries used to extract relevant studies for each topic.

Topic	Search query
Clinical NLP	("coreference" OR ("semantic textual similarity" OR STS) OR ("named entity recognition" OR NER) OR "relation extraction" OR "natural language inference" OR "question answering" OR "entity normalization") AND (BERT OR Transformer) AND ("clinical" OR "medical" OR "biomedical" OR "EHR") from 2017
Structured EHR	(Transformer OR BERT) AND ("deep learning" OR "machine learning") AND (EHR OR "electronic health records") from 2017
Medical Imaging	(Segmentation OR registration OR "image captioning" OR "report generation" OR "visual question answering" OR "image synthesis" OR "classification" OR "reconstruction") AND ("Transformer" OR "vision transformer") AND ("clinical" OR "medical" OR "biomedical" OR "EHR") from 2017
Critical Care	(Transformer) AND ("deep learning" OR "machine learning") AND ("critical care" OR "surgery" OR "surgical") from 2017
Social Media	(Transformer OR BERT) AND ("deep learning" OR "machine learning") AND ("social media" OR "crowdsourcing" OR "crowdsourcing" OR "twitter" OR "tweet") from 2017
Bio-physical Signals	(Transformer OR BERT) AND ("deep learning" OR "machine learning") AND ("medical" OR "health" OR "clinical" OR "biomedical") AND ("signal" OR "ECG" OR "EMG" OR "EEG" OR "human activity" OR "HAR") from 2017
Biomolecular Sequences	(Transformer OR BERT) AND ("deep learning" OR "machine learning") AND (DNA OR RNA OR gene OR genome OR genomic OR transcriptomic OR protein OR proteomic OR metabolite OR metabolism OR metabolomic OR chromosome OR receptor OR mitochondria OR splicing) from 2017

one reviewer. Six reviewers worked independently on studies pertaining to different domains. For topics which had a significantly larger number of papers (for e.g., clinical NLP and medical imaging) three reviewers worked together to analyze relevant articles, and only those articles were retained which were deemed relevant by all three reviewers.

We identified the top keywords from articles included in this report to provide an overview of key concepts, data modalities, and tasks. The word cloud in Fig. 2 shows the 50 most common keywords across articles, with a larger font representing more papers; while Fig. 3 shows data modalities and the corresponding tasks.

## 3. Background

Transformers are multilayered neural networks formed by stacking either multiple encoder and/or decoder blocks which utilize the attention mechanism, as explained in the following section.

### 3.1. Attention

The attention mechanism computes similarity between individual input tokens, such as vectors of word embeddings. In a basic Transformer architecture, each input embedding generally can take three roles: (1) Query  $Q$  which is the current focus of the attention mechanism and is being compared to all other input tokens, (2) Key  $K$  is the input token being compared to the query, and (3) Value  $V$  is a value used to compute the output of attention. The attention function can be considered a mapping between a query and a set of key-value pairs to produce an output [2].

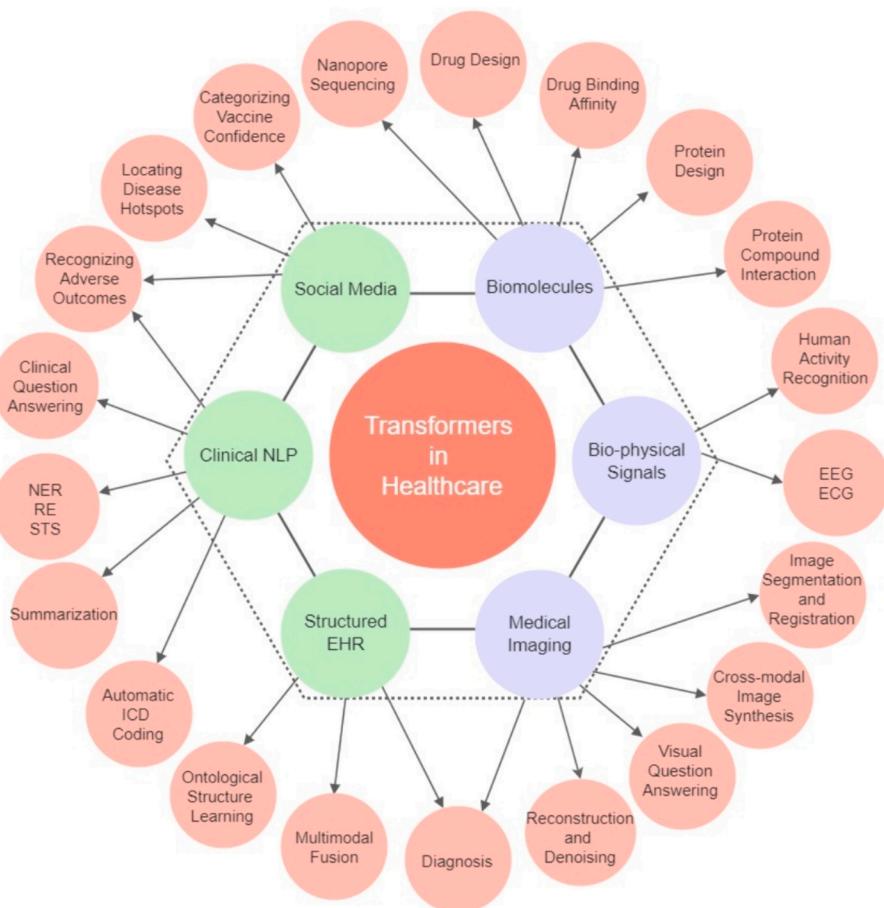
We will represent the input  $X \in R^{n \times d}$  as a sequence of  $n$  tokens with an embedding dimension of  $d$ . The input sequence  $X$  is linearly transformed into query  $Q$ , key  $K$ , and value  $V$  using Eqs. (1), (2), and (3), respectively.

$$Q = X \bullet W_q \quad (1)$$

$$K = X \bullet W_k \quad (2)$$



**Fig. 2.** Word cloud depiction of keywords used in the surveyed literature. Abbreviations. BERT; Bidirectional Encoder Representations from Transformers, CNN; Convolutional Neural Networks, EHR; Electronic Health Records, MRI; Magnetic Resonance Imaging, NER; Named Entity Recognition, NLP; Natural Language Processing, STS; Semantic Textual Similarity.



**Fig. 3.** Major healthcare data modalities and corresponding tasks. Abbreviations: EEG; Electroencephalography, ECG; Electrocardiogram, NER; Named Entity Recognition, RE; Relation Extraction, STS; Semantic Textual Similarity, ICD; International Classification of Diseases, EHR; Electronic Health Record.

$$V = X \bullet W_v \quad (3)$$

where  $W_q$ ,  $W_k$ , and  $W_v$  are the weight matrices to obtain  $Q$ ,  $K$ , and  $V$  matrices. The  $Q$ ,  $K$ , and  $V$  are then used to compute the scaled dot product attention as shown in Equation.

$$\text{Attention}(Q, K, V) = \text{softmax}(\alpha Q \bullet K^T) \bullet V \quad (4)$$

In Eq. (4), the scaled dot product operation is performed between the query and key matrices, followed by a softmax function. Here, the scale factor  $\alpha$  is used to mitigate the vanishing gradient problem or numerical instability. It is typically chosen to be  $1/(\sqrt{d_k})$ , where  $d_k$  is the key dimension.

### 3.2. Attention mechanisms

Transformer models primarily use three types of attention: self-attention, masked self-attention, and cross-attention.

#### 3.2.1. Self-attention

Self-attention is when attention is computed between tokens in the same sequence. The self-attention block is found in the Transformer encoder. The dimensions of query, key, and value are the same in self-attention, i.e.,  $d_k = d_q = d_v$ .

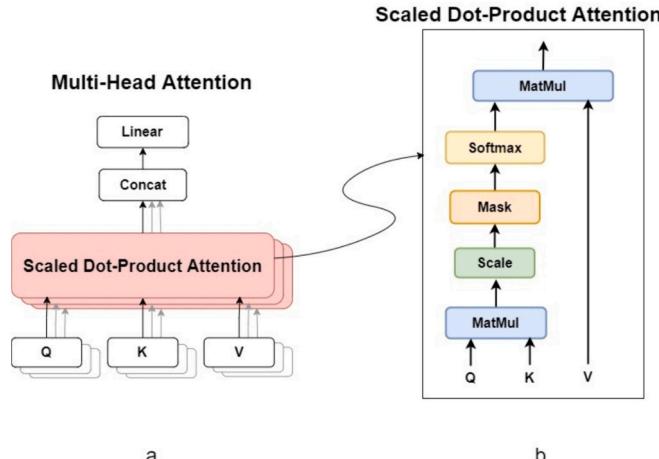
#### 3.2.2. Masked self-attention

In sequence prediction problems, such as machine translation, the context of previous tokens in a sequence is used to predict the subsequent output. A mask is typically employed to prevent the model from attending to subsequent tokens in a sequence. The mask  $M$  is a square upper triangular matrix with dimension  $n$ , where  $n$  is the number of tokens in the input sequence. The mask is applied to the scaled dot product of the query and key via element-wise addition, as in Eq. (6)

$$\text{Masked Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \bullet K^T}{\sqrt{d_k}} + M\right) V \quad (6)$$

#### 3.2.3. Cross-attention

Cross-attention is attention computed between tokens of one sequence with tokens of another sequence. In Transformer, the input and output sequences interact through cross-attention in the decoder module. The cross-attention module receives its queries from the previous masked self-attention layer of the decoder and its keys and values from the last encoder. Queries correspond to the desired output sequence, while the keys and values are generated based on the input sequence in the encoder.



**Fig. 4.** Multi-head attention mechanism. In the encoder and decoder, multiple attention heads are stacked together and their outputs are concatenated.

#### 3.2.4. Multi-head attention

It has been shown that compared to a single attention computation, multiple attention operations can improve the model's performance by capturing different similarity relationships in the sequence [2]. The attention blocks in both the encoder and decoder are computed with  $h$  attention heads, as shown in Fig. 4. The original Transformer model employed  $h = 8$  attention heads. Every attention head has three learnable weight matrices:  $W_q^i$ ,  $W_k^i$ , and  $W_v^i$ , where  $i$  represents a particular attention head. The attention outputs from multiple heads (denoted by  $\mathcal{H}$ ) are then concatenated and linearly transformed to the model dimension with a parameter matrix  $W_o$ .

### 3.3. Position-wise feed-forward network

The output of the attention modules is passed to a two-layered feedforward network (FFN). The FFN performs an independent position-wise linear transformation on each token of the sequence. Parameters of this network are shared across all positions of the sequence.

Let  $\mathcal{H}$  be the output of the multi-head attention block and  $d$  be the model dimension. The first linear layer transforms  $\mathcal{H}$  from dimension  $d$  to an intermediate dimension  $d_f$ , also referred to as the feedforward dimension. The second linear layer transforms the output of the first linear layer from  $d_f$  to the original model dimension  $d$ . The FFN is given by Eq. (9).

$$\mathcal{T}(H) = \text{ReLU}(\mathcal{H} \bullet W_1 + b_1) \bullet W_2 + b_2 \quad (9)$$

The intermediate dimension  $d_f$  is usually set to a value larger than  $d$ .

### 3.4. Residual connections and layer normalization

Residual connections [17] allow gradients to skip non-linear activation functions, followed by layer normalization. Layer normalization scales the values of all hidden layers to a similar range to avoid exploding or diminishing values obtained through a chain of multiplication operations.

### 3.5. Positional encodings

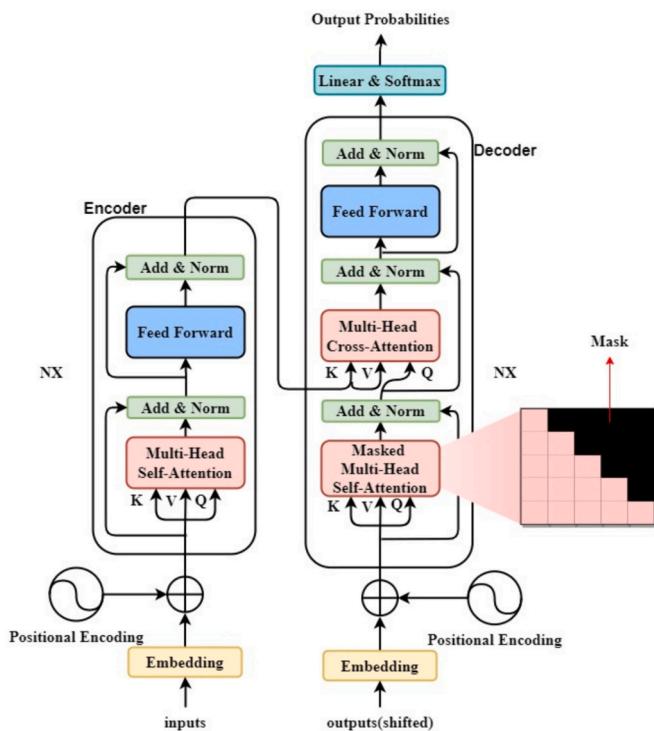
Because the self-attention module attends to all tokens of a sequence in parallel, it intrinsically neglects the order of tokens in the sequence. This necessitates using a positional encoding (PE) vector that denotes the unique position of each token. Transformers use a combination of sine and cosine functions of different frequencies to create PE vectors shown in Eq. (10). PE vectors are added to the embeddings of each input token; therefore the PE dimension is chosen to be the same as the embedding dimension. Since sine and cosine functions have values in the range [-1, 1], the values of the positional encoding matrix are constrained to a normalized range. This technique enables Transformers to capture the relationship between items that are both close and far from one another in a sequence.

$$PE_{(pos,i)} = \begin{cases} \sin(pos \bullet \omega_k), & \text{if } i = 2k \\ \cos(pos \bullet \omega_k), & \text{if } i = 2k + 1 \end{cases} \quad (10)$$

$$\omega_k = \frac{1}{2k/d}, k = 1, 2, \dots, \frac{d}{2} \\ 10000$$

### 3.6. Assembling a transformer

Transformer consists of an encoder and a decoder network. The encoder consists of identical encoder blocks stacked upon each other, each consisting of a self-attention and an FFN layer. The decoder consists of stacked identical decoder blocks, each consisting of a masked self-attention layer, cross-attention layer, and FFN layer. The encoder transforms an input sequence into encoded representations, while the



**Fig. 5.** Schematic of the transformer architecture [2,18].

decoder operates upon these representations.

The original Transformer architecture (Vaswani et al., 2017) [2], shown in Fig. 5, has six identical stacked encoders and six identical stacked decoder blocks. Each encoder block comprise of multi-head self-attention followed by FFN. Every decoder block consists of multi-head masked self-attention, multi-head cross-attention, and FFN arranged sequentially. The cross-attention layers attend to queries from the previous masked attention layers, whereas keys and values are obtained from the output of the final encoder block. The output of the last encoder is used to obtain the keys and values to compute the multi-head cross attention in all the decoder layers.

### 3.7. Computational complexity of transformer attention

The self-attention mechanism of Transformer can attend to variable-length input size but has  $O(n^2 \cdot d)$  time complexity where  $n$  and  $d$  are the input sequence length and the model dimension. For long input sequences, this attention computation becomes computationally expensive. Many Transformer variants try to reduce the computational complexity via different approaches [19].

### 3.8. Transformer model usage

In general, Transformer architectures can be divided into three categories.

**Encoder-Decoder:** consists of multiple encoder and decoder blocks and is typically used in sequence-to-sequence modeling tasks, such as machine translation.

**Encoder only:** Only the encoder blocks are used to model the input sequence. The output of the encoder is a contextual representation of the input sequence. This type of architecture is used for classification or label prediction problems (most models in this review).

**Decoder only:** Only decoder blocks are used. This architecture is used for sequence generation, image captioning, and language modeling tasks.

## 4. Mainstream transformer-based architectures

In this section, we will discuss the two prominent transformer-based architectures with significant impact on NLP and computer vision.

### 4.1. Bidirectional encoder representations from transformers (BERT)

BERT [20], is an encoder-only Transformer architecture that can produce rich contextualized word and sentence embeddings for NLP. Unlike traditional language models, which read text input sequentially (left-to-right or right-to-left), the Transformer encoder in BERT reads the entire sequence of words at once, thereby learning a richer representation of context and information flow in a sentence. The BERT architecture uses self-supervised pretraining steps, namely Masked Language Modeling (MLM), to create context-sensitive word embeddings, and Next Sentence Prediction (NSP) to model sequential association between sentences. MLM masks a fraction of the input tokens and aims to predict them based on their context. This helps to disentangle ambiguity in the text by using surrounding text to establish context. In NSP, a combination of two sentences is fed to the Transformer encoder. In 50% of cases, the second sentence is the next sentence in the original text, while in the remaining 50% of cases, the second sentence is randomly selected. The encoder learns to distinguish scenarios where the sentences are logically linked. When training the BERT model, MLM and NSP are trained together to minimize the combined loss function of the two strategies. BERT can be used for various language tasks, such as sentence classification, Question Answering (QA), and Named Entity Recognition (NER) with finetuning and minor modifications to the original architecture.

### 4.2. Vision Transformer (ViT)

ViT is a pure Transformer architecture without convolutional layers and was proposed for image classification tasks [1]. Like BERT, ViT is also an encoder-only Transformer model. Transformers cannot directly process spatial data such as images; therefore, data must be converted to a sequence. ViT splits an image into fixed-size patches, generally  $16 \times 16$  or  $32 \times 32$  flattened, before they are provided as an input to the transformer model. The flattened patches are placed in a sequence, then transformed into a low-dimensional linear embedding. Like the original Transformer, PEs are added to the linear embeddings to inject information about each patch's relative location in the image, where 1D, 2D, and learnable positional embeddings can be used. An extra learnable class embedding is added at the start of the sequence, used for downstream classification tasks. During fine-tuning, a classification head comprised of a single hidden layer network is attached to this class embedding.

Transformer models by design do not possess the inductive biases of CNNs, such as limited receptive field and translational invariance (ability to detect or recognize an object regardless of its location in an image). In CNNs, the receptive field increases linearly with the depth of the model. While the Transformer lacks the inductive biases of the CNN, they are permutation invariant (not dependent on the order of elements in a sequence), and the shallow layers of the model can attend to the entire image.

## 5. Large language models (LLMs)

Foundation models are large-scale AI systems trained on vast amounts of data to be adapted for a wide range of downstream tasks [21]. LLMs colloquially refer to a class of foundation models with billions of parameters trained on language corpora with billions of words to generate human-like language and solve different NLP tasks. Most LLMs use the Transformer architecture, the current default architecture for processing sequential data as of 2023. The success of LLMs comes from the self-supervised pre-training paradigm, which takes advantage of large free text data without annotation. This pre-training technique

enabled LLMs to generate coherent and realistic language, making them useful for various applications such as text completion, dialogue generation, and content generation. BERT style LLMs (Encoder-Decoder or Encoder only) are pretrained using masked language modeling while GPT style (decoder only) models are pretrained by generating next word in a sequence. Generative AI models trained to generate text and question answering tasks are autoregressive decoder-only language models. Examples of autoregressive decoder-only language models include PaLM [22], GPT-3 [23], Chinchilla, LLaMA [24], PaLM2 [25] used in BARD chatbot, and GPT-4 [26]. These models are trained on billions of tokens obtained from datasets such as Common Crawl, WebText2, Books1, Books2, Wikipedia, Stack Exchange, PubMed, ArXiv, GitHub, Gutenberg, and many more. Some of the domain-specific LLMs include Galactica [27], trained on curated human scientific knowledge corpora, BloombergGPT [28], trained on proprietary financial data, and CodeX [29] for code generation. A timeline of popular LLMs is displayed in Fig. 6.

The number of parameters in LLMs and the size of their training data has increased rapidly, reaching up to trillions of tokens [24]. The capabilities of LLMs appear to be a function of the amount of data, parameters, and computation resources rather than architectural design advancements [30]. The scaled-up language models develop abilities beyond the trained outcomes called ‘emergent abilities,’ which are not designed but discovered after deployment [31]. For example, GPT-3 showed few-shot prompting ability; when provided few input-outputs for a natural language task, the model can perform the task on unseen samples without further training or gradient updates to the parameters [23]. Parameter-efficient models such as Stanford Alpaca [32] and efficient finetuning approaches of Quantized LLMs such as QLoRA [32] have been introduced to address situations where computational resources are limited. Despite the exceptional ability of LLMs to generate realistic text, they can also generate false information, toxic language, and racial stereotypes [33,34].

In the medical domain, Agrawal et al. [35] demonstrated that LLMs can be few-shot clinical information extractors without further training on the clinical data. They used InstructGPT [36] for this task, significantly outperforming existing zero-shot and few-shot baselines. In Radiology, Jeblick et al. [37] performed an exploratory case study to evaluate ChatGPT’s ability to simplify radiology reports. Expert human radiologists considered the simplified reports complete, factual, and devoid of harmful text that could misguide the patient. However, instances of missing key findings and incorrect statements were observed. The PMC-LLaMA [38] model, fine-tuned on 4.8 million biomedical papers obtained from PubMed Central, demonstrated a better understanding of biomedical domain-specific concepts than the original LLaMA when evaluated on biomedical QA benchmarks. GatorTron [39], a large clinical language model with 8.9 billion parameters trained on over 90 billion words of clinical text, was applied to clinical NLP tasks such as clinical concept extraction. Luo et al. [40] proposed BioGPT, a biomedical domain specific generative model pretrained on PubMed abstract corpus to generate fluent biomedical term descriptions.

Singhal et al. [41] evaluated the 540 billion parameters PaLM [22] and its variant FLAN-PaLM [42] on the benchmark dataset MultiMedQA.

This benchmark dataset combines multiple QA datasets, including medical exams, consumer queries, and research. The authors also introduced Med-PaLM, a parameter-efficient model that used prompt instruction tuning to fix the critical Flan-PaLM gaps observed upon human evaluation. In subsequent work, Singhal et al. proposed Med-PaLM2 [43] to bridge the gap between the model’s answers to that of clinicians. The model combines improvements that come with PaLM2 [25], a novel ensemble refinement prompting strategy, and domain-specific model fine-tuning. Scaled-up models such as ChatGPT, PaLM, PaLM2, and GPT-4 have been shown to answer medical questions and successfully pass or achieve near-passing scores on medical licensing examinations [41,44–47]. These existing large medical foundation models trained on broad biomedical domain corpora such as PubMed are tested on tasks with minimal significance to the health systems [48]. The impressive advancements of foundation models have not yet permeated into medical AI. These early approaches are limited by a lack of large, diverse medical datasets, the complex nature of medical data, federal patient data privacy regulations, and the recency of the general-purpose foundation models [49]. Transformers in NLP

## 6. Transformers in clinical NLP

### 6.1. Clinical word embeddings

Word embeddings map variable-length words to a fixed-length vector while preserving syntactic and semantic information. Word embeddings are a standard representation used in NLP. Traditional word embedding techniques such as word2vec [50] or GloVe [51] learn an aggregated representation of all contexts associated with a word. Previously contextual word embedding based on models such as ELMo [52], BERT [20], and ULMFiT [53] achieved state of the art (SOTA) performance on NLP tasks. However, these embeddings cannot be adapted directly to clinical or biomedical text due to differences in the linguistic domain corpora. Lee et al. [50] introduced BioBERT, a pre-trained language model in the biomedical domain, to overcome this difficulty. BioBERT is initialized with BERT weights and is pre-trained on PubMed Central full-text articles and abstracts as shown in Fig. 7. This pre-trained model is fine-tuned on three popular biomedical NLP tasks: Named Entity Recognition (NER), Relation Extraction (RE), and QA. BioBERT has outperformed previous models on biomedical text mining tasks with minimal task-specific modification.

Further specialization of BERT and BioBERT via pre-training on specific EHR databases has proven promising. Alsentzer et al. [55] pre-trained BERT and BioBERT on 2 million clinical notes from the MIMIC-III database [56] to obtain clinical BERT and Bio+Clinical BERT. Si et al. [57] explored various embedding methods such as word2vec [50], GloVe [51], fastText [58], ELMo [52], and BERT [20] on clinical concept extraction tasks to demonstrate the generalizability of these traditional embedding methods. When pre-trained on a clinical domain-specific corpus [56], all the embeddings yielded increased performance. Huang et al. [59] pretrained BERT [20] on clinical notes from the MIMIC-III dataset [56] to develop ClinicalBERT. ClinicalBERT achieved

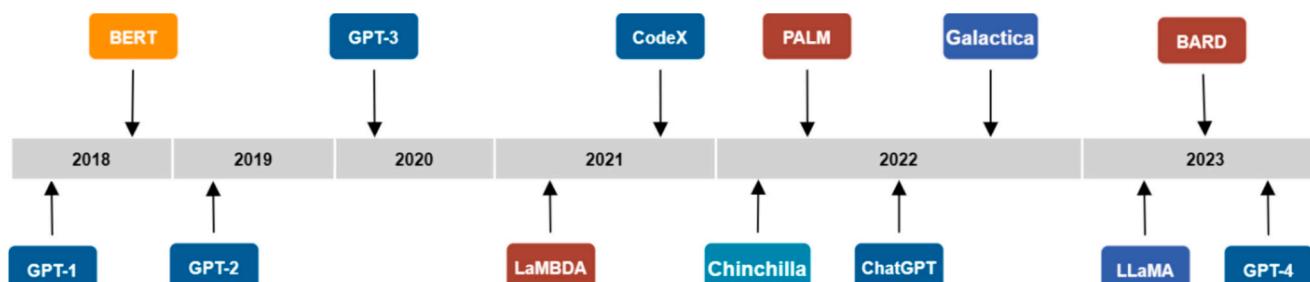
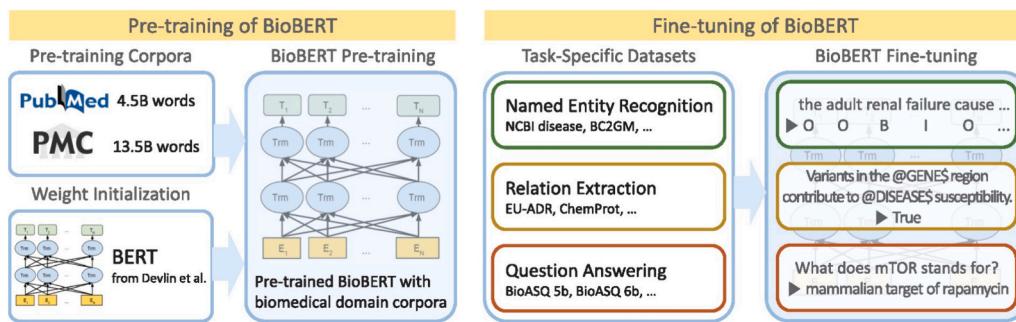


Fig. 6. The timeline of popular large language models developed over the years (2018-2023).



**Fig. 7.** BioBERT pre-training and finetuning overview. Source: Image adapted from [54] without modifications.

higher Pearson correlation scores than word2vec [50] and fastText [58]. All these models were pre-trained on clinical domain corpora and have outperformed models pre-trained on general or biomedical domain corpora in clinical NLP tasks.

## 6.2. Transformers for clinical information extraction (IE)

EHRs contain a wealth of patient information stored in structured and unstructured formats, including detailed clinical notes used for documentation. Parsing through this data is difficult due to the unstructured nature of the free text entries recorded by clinical staff in the EHR. Clinical IE consists of sub-tasks such as NER, coreference resolution (CR), QA, semantic textual similarity (STS), relation extraction (RE), and entity normalization (EN). The success of Transformers inspired researchers to adapt Transformer-based architectures for clinical IE (Table 2).

### 6.2.1. Named entity recognition

Clinical named entity recognition (CNER) aims to identify entities, concepts, and events such as diseases, drugs, treatments, medical conditions, and symptoms from clinical narratives. CNER is challenging as clinicians often use acronyms and abbreviations to describe complex clinical terms without using standardized clinical ontology. Earlier approaches used the BERT model to generate clinical textual embeddings, which were further used to train other deep learning models, such as Bi-LSTM and conditional random fields [114–116]. Later, for biomedical and clinical domains, domain-specific BERT-based models such as BioBERT [54] and clinical BERT by Alsentzer et al. [55] established baselines on CNER datasets. BERT-based models have been applied to CNER tasks in different languages, such as Chinese [117,118], Korean [119], Italian [120], Spanish [121], and Arabic [122].

The clinical de-identification task, which removes protected health information, was also approached as a NER problem by pretrained BERT-based models, such as clinical-BERT [55] and UMLS-BERT [123]. These models were applied to i2b2-2006 [75] and i2b2-2014 [77] de-identification tasks. Garcia et al. [124] and Mao et al. [125] used BERT on the MEDDOCAN [126] Spanish de-identification corpus.

The clinical concept extraction task predicts a concept's start and end positions in a document. BIO tags are commonly used, where "B", "I", and "O" refer to the beginning, inside, and outside of a concept. Yang et al. [92] developed an open-source Transformers package with four transformer-based models, BERT [20], ALBERT [96], RoBERTa [95], and ELECTRA [97], pretrained on MIMIC-III dataset for clinical concept extraction. Peng et al. [83] used transfer learning to fine-tune BERT [20] for concept extraction on BC5CDR [62] and ShARe/CLEF [110] datasets. Khan et al. [127] proposed MT-BioNER, a transformer-based model for intent classification and slot tagging. The authors combined BERT encoder layers with task-specific layers to train their model on NCBI-disease [128], BC5CDR [62], and JNLPBA [89] datasets.

### 6.2.2. Clinical coreference resolution (CR)

The CR task aims to identify all mentions of the same entity in a text. Trieu et al. [129] performed CR in full-text articles as part of the CRAFT 2019 shared task [130]. The authors employed a span-based end-to-model proposed by Lee et al. [131] and replaced the LSTM layers with BERT. Their results on the CRAFT coreference resolution task indicate the effectiveness of BERT in capturing long-distance coreferences in large documents. Steinkamp et al. [132] used BERT [20] to perform CR for symptom extraction on the i2b2 2009 Medication Challenge [133] and MIMIC-III datasets [56], showing better performance compared to recurrent models.

### 6.2.3. Clinical relationship extraction (CRE)

CRE is categorized into concept relationship and temporal relationship extraction. Concept relationship extraction identifies the relationship between two concepts (e.g., drug and dosage), whereas temporal relationship extraction evaluates the relationship between clinical events occurring at different times. Peng et al. [83] approached the CRE task as a sentence classification problem by replacing named entity mentions of interest with pre-defined tags using BERT [20] on DDI [86], ChemProt [70], and i2b2 2010 [61] datasets. Wei et al. [98] fine-tuned BERT outperformed SOTA RE models on clinical RE tasks using n2c2-2018 [94] and i2b2-2010 [61] datasets. Zhang et al. [114] pretrained the BERT model on Chinese clinical text and fine-tuned on the breast cancer dataset to classify the relationship between clinical concepts and corresponding attributes for breast cancer. Using BERT, Xue et al. [129] used an integrated joint learning approach for NER and CRE in coronary angiography Chinese clinical text. Lai et al. [134] proposed BERT-GT, which combines BERT with Graph Transformer by integrating the neighbor attention mechanism into BERT. BERT-GT was used for cross-sentence RE on the N-ary [135] and BioCreative CDR [136] datasets. Lin et al. [137] developed a pre-trained BERT model on the MIMIC-III dataset and BioBERT [54] models for temporal RE on the THYME [138] corpus. Their BioBERT model with sentence agnostic 60-token window approach was used for the CONTAINS temporal relation extraction task on the colon cancer test set.

### 6.2.4. Question answering (QA)

The QA ability of a model can serve as an indicator of its ability to learn the medical text. Jin et al. [90] introduced the PubMedQA dataset for biomedical research question answering, and fine-tuned BioBERT model to establish a baseline on the dataset. Yoon et al. [106] pretrained the BioBERT model on SQuAD [107,108] datasets and fine-tuned it for the BioASQ [71,91] biomedical QA challenge. This model achieved SOTA performance on factoid, list, and yes/no type questions of the BioASQ dataset. He et al. [139] proposed a procedure for consumer health question answering and medical language inference tasks using models such as BERT[20], BioBERT[54], SciBERT[100], ClinicalBERT [55], BlueBERT[83], and ALBERT[96]. Schmidt et al. [140] developed a QA-BERT model for question answering using the PICO (Population, Intervention, Comparator, and Outcome) framework. The PICO element

**Table 2**  
Transformers in clinical NLP.

Reference	Title	Tasks	Datasets	Architecture
Lee et al. [54]	BioBERT: a pre-trained biomedical language representation model for biomedical text mining	NER, RE, QA	NCBI Disease [60], i2b2 2010 [61], BC5CDR [62], BC4CHEMD [63], BC2GM [64], JNLPBA [65], LINNAEUS [66], Species-800 [67], GAD [68], EU-ADR [69], CHEMPROT [70], BioASQ [71]	BERT[20]
Alsentzer et al. [55]	Publicly available clinical BERT embeddings	NLI, NER, de-identification, concept extraction, entity extraction	MIMIC-III [56], i2b2 2010 [61], i2b2 2012 [72,73], MedNLI [74], i2b2 2006 [75], i2b2 2014 [76,77]	BERT [20]
Si et al. [57]	Enhancing clinical concept extraction with contextual embeddings	Concept extraction	i2b2 2010 [61], i2b2 2012 [72], i2b2 2014 [76], ShARe/CLEF [78,79], SemEval [80–82], MIMIC-III [56]	BERT [20]
Peng et al. [83]	BlueBERT: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets	SS, NER, RE, DC, Inference	MEDSTS [84], BIOSSES [85], BC5CDR [62], ShARe/CLEF [78], DDI [86], CHEMPROT [70], i2b2 2010 [61], HoC [87], MedNLI [74]	BERT[20]
Gu et al. [88]	Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing	NER, RE, SS, DC, QA	NCBI Disease [60], BC5CDR [62], BC2GM [64], JNLPBA [89], CHEMPROT [70], DDI [86], GAD [68], BIOSSES [85], HoC [87], PubMedQA [90], BioASQ [71,91]	PubMedBERT
Huang et al. [59]	ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission	Patient readmission prediction	MIMIC-III [56]	BERT [20]

**Table 2 (continued)**

Reference	Title	Tasks	Datasets	Architecture
Yang et al. [92]	Clinical concept extraction using transformers	Concept extraction	MIMIC-III [56], i2b2 2010 [61], i2b2 2012 [72,73], n2c2 2018 [93,94]	BERT [20], RoBERTa [95], ALBERT [96], ELECTRA [97]
Wei et al. [98]	Relation Extraction from Clinical Narratives Using Pre-trained Language Models	RE	n2c2 2018 [93,94], i2b2 2010 [61]	BERT [20]
Mayer et al. [99]	Transformer-Based Argument Mining for Healthcare Applications	Argument component detection, RE	MEDLINE	BERT [20], BioBERT [54], SciBERT [100], RoBERTa [95]
Huang et al. [101]	Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation	Prognosis prediction	MIMIC III [56]	XLNet [102], BERT [20], ClinicalBERT [59]
Yu et al. [103]	BioBERT based named entity recognition in electronic medical record	NER	I2b2 2010 [61]	BioBERT[54]
Alimova et al. [104]	Multiple features for clinical relation extraction: A machine learning approach	RE	n2c2 2018 [93,94], MADE 2018 [105]	BERT [20], BioBERT [54], ClinicalBERT [59]
Jin et al. [90]	PubMedQA: A dataset for biomedical research question answering	QA	PubMedQA [90]	BioBERT [54]
Yoon et al. [106]	Pre-trained language model for biomedical question answering	QA	SQuAD [107,108], BioASQ [71,91]	BioBERT [54]
Ji et al. [109]	BERT-based ranking for biomedical entity normalization	EN	ShARe/CLEF [110], NCBI [60], TAC2017ADR [111]	BERT [20], BioBERT [54], ClinicalBERT [55]
Yang et al. [112]	Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models	STS	2019 n2c2/ Open Health NLP [113]	BERT [20], XLnet[102], RoBERTa [95]

NER: Named Entity Recognition; SS: Sentence Similarity; RE: Relation Extraction; DC: Document Classification; NLI: Natural Language Inference; QA: Question Answering, EN: Entity Normalization; STS: Semantic Textual Similarity

dataset [141] was combined with SQuAD datasets [107,108] to increase the generalizability and flexibility of the model on all types of questions. The proposed QA-BERT performed better than LSTM and BERT baselines [140].

#### 6.2.5. Biomedical entity normalization (BEN)

BEN aims to link mentions of an entity in a clinical document (e.g., EHR) to their corresponding concepts in a knowledge base [142]. Ji et al. [109] fine-tuned pre-trained models such as BERT [20], BioBERT [54], ClinicalBERT [55] on three different datasets ShARe/CLEF [110], NCBI [60], TAC2017ADR[111] for performing BEN. Li et al. [143] proposed the EhrBERT model, pre-trained on 1.5 million EHR notes, and evaluated it on three entity normalization corpora, namely the MADE corpus [105], NCBI disease corpus [60], and CDR corpus [62]. Authors observed that their models performed worse when the pre-training domain and fine-tuning task were distant.

#### 6.2.6. Semantic text similarity (STS)

STS is an NLP task that measures the similarity between two pieces of text using a pre-defined metric. Xiong et al. [144] proposed a gated network to fuse one hot and distributed representations obtained from sentence-level features like inverse document frequency, sentence length, N-gram overlaps, and similarity metrics between two input sentences. Their fusion-gated BERT model was used on the clinical STS task of the BioCreative/OHNLP 2018 challenge [145]. Yang et al. [112] explored three models, BERT [20], XLnet [102], and RoBERTa [95], for clinical STS as a part of the 2019 n2c2/Open Health NLP challenge [113]. The models were pre-trained on a general STS dataset and fine-tuned on the clinical STS training partition. Among these, RoBERTa-large achieved the highest performance.

#### 6.2.7. Automatic international statistical classification of diseases (ICD) coding

ICD codes are a set of alphanumeric designations to communicate diseases, symptoms, procedures, diagnoses, and abnormal findings in a universally accepted way among healthcare professionals. ICD coding involves recording the ICD codes associated with a patient's visit. This coding process is often performed manually, which may result in documentation errors and consume a significant amount of time. Zhang et al. [146] proposed BERT-XML with multi-label attention to model 2292 ICD-10 codes from EHR notes [147]. Biswas et al. [148] used a transformer-based encoder architecture TransICD with a structured self-attention mechanism [149] to extract label-specific representations for multi-label ICD coding. Label distribution aware margin loss [150] was used to address the imbalance in ICD codes data. Transformer-based automatic ICD coding was used in clinical texts of Chinese [151], Spanish [152,153], Swedish [154], and Thai [155]. Silvestri et al. [156] used a Transformer Cross-lingual Language Model(XLM) [157] for automatic ICD coding by fine-tuning clinical texts in English and testing on clinical Italian text.

#### 6.2.8. Neural machine translation (NMT)

Automatic NMT of biomedical data is essential to make essential healthcare information available to healthcare professionals overcoming language barriers. Tubay et al. [158] for the low-resourced biomedical NMT task used a Transformer model enhanced with a multi-source translation technique capable of exploiting multiple text inputs from the same language family. Berard et al. [159] proposed a multilingual neural machine translation (MNMT) model to translate biomedical text from 5 different languages (French, Spanish, German, Italian, and Korean) to English. The MNMT model is a variant of Transformer Big architecture with complex encoder capable of representing multiple languages. Liu et al. [160] proposed BioNMT Transformer model to translate domain specific biomedical vocabulary from foreign languages. The model is capable of semantic disambiguation of unknown words in the translation using external biomedical

ictionaries. Wang et al. [161] used a Transformer large model with 20 encoder layers for biomedical translation shared task to translate German, French, and Spanish to English. Subramanian et al. [162] used a Transformer model for the same biomedical shared task at WMT to translate text from English to German and Russian. Their transformer model used a combination of model scaling, data augmentation with back-translation, knowledge distillation, model ensembling, and noisy channel re-ranking to perform the translation task.

### 7. Transformers for structured EHR data

Structured EHR data includes ICD codes for diagnoses, medication, vital signs, laboratory tests and other demographics collected every time a patient visits the hospital. These data are linked by an underlying temporal structure representing the cycle of diagnosis, medication/intervention, and potential patient readmissions. Furthermore, medication and diagnosis codes are derived from an ontological tree structure. Therefore, clinical tasks such as predicting future disease diagnoses, readmissions, or mortality rely on accurately representing the temporal and graphical structure of a patient's EHRs. This challenge has led to three broad NLP tasks on structured EHR content that have been attempted in recent years using transformer networks.

#### 7.1. Ontological structure learning

Previous studies have tried to learn the graphical structure inherent within the EHR using novel Transformer architectures. Choi et al. proposed the Graph Convolution Transformer (GCT) to jointly learn the relationships between diagnoses and medication codes while performing diagnosis-treatment classification [163]. They used conditional probabilities between medications and diagnoses calculated over the entire dataset to guide the attention maps in their Transformer network. Their model was validated on the eICU collaborative research dataset [164]. In contrast, Shang et al., 2019 explicitly used graph neural networks (GNN) for learning medical ontology embeddings and used these embeddings in a transformer to recommend future medications using the MIMIC-III dataset [165]. To leverage the entire dataset, they pre-trained G-BERT, a combination of GNN and BERT, on EHR data with only one admission. Peng et al., used a graph-based attention model (GRAM) to create ontological embeddings, which were then represented using multi-head self-attention to learn the ontological structure of medications within EHR [166].

#### 7.2. Multi-modal data fusion

Previous studies have used Transformer networks to create joint embeddings amongst multiple data modalities, such as EHR and clinical notes. Darabi et al., used separate Transformer networks to create different representations for the clinical codes (ICD, drug, and procedure) and clinical notes and combined them into one "patient representation" [167]. They used this joint representation to predict future diagnoses, procedures, length of stay (LOS), readmission, and mortality. Studies have used joint-embeddings in BERT to predict rare diseases such as chronic cough [168] or depression [169]. Xu et al., proposed the use of multi-modal fusion architecture search (MUFASA), using an evolutionary algorithm to jointly search for the optimal architecture to represent subsets of EHR data and the optimal stage at which the individual embeddings will undergo fusion [170]. In contrast, Zhang et al., used a contrastive learning approach to increase the mutual agreement between different modalities for the same patient and increase the contrast for the same modality amongst different patients while jointly optimizing a prediction loss [171]. They showed that combining this representation with the BERT encoder predicted mortality and length of stay better than other baselines.

### 7.3. Predicting future diagnoses using ICD codes

BEHRT, an adaptation of BERT on EHR data, was trained from scratch using the masked language modeling task on sequential ICD codes and age to predict future diagnoses [172]. This model was developed primarily on the UK Clinical Practice and Research Datalink (CPRD) [173]. Recently, BEHRT was used to predict incident heart failure [174] and to perform causal inference [175]. The Hi-BEHRT model extended this by incorporating self-supervised pretraining by masking certain EHR data and certain time points in patients' visitation history and creating localized feature aggregator Transformer embeddings fused at a later stage using global attention [176]. Hi-BEHRT performed better than BEHRT in predicting the onset of heart failure, diabetes, chronic kidney disease, and stroke. Compared to the BEHRT-based models, Med-BERT expanded the pretraining task to include prediction of prolonged length of stay and used a combination of ICD-9 and ICD-10 codes to create their model, which was subsequently evaluated on predicting diabetes and heart failure [177]. Another model, HiTANet [178], explicitly included a time vector to represent the time elapsed between consecutive visits. The time embedding was combined with the original visit embedding and used as key values in a global attention block to represent the most significant time points in a patient's medical history. They tested their model efficacy in predicting future diagnoses of three disease-specific datasets. The RAPT model combined an explicit time-span information vector with additional pre-training tasks such as similarity prediction and reasonability check to address data insufficiency, incompleteness, and short sequence problems inherent in EHR data [179]. They evaluated their model for predicting pregnancy outcomes, risk period, and the diagnoses of diabetes and hypertension during pregnancy.

## 8. Transformers in medical imaging

### 8.1. Medical image segmentation

Image segmentation is a dense pixel classification task that captures the complex interactions between individual pixels of an image. Unlike general-purpose image segmentation, medical image segmentation suffers from a lack of large datasets, requires the context of surrounding anatomical structures, and must account for inter-patient anatomical variabilities. Several data modalities, such as X-ray, Ultrasound, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), and microscopy, can benefit from medical image segmentation. Before Transformers, the U-net architecture, proposed by Ronneberger et al. [180], was the prominent architecture for medical image segmentation. The U-net model is a Convolutional Neural Network (CNN). Convolutional layers are limited in long-range feature modeling. This is because the receptive field of convolutional filters increases linearly; therefore, only the deepest convolutional layers have the global context of an image. Although incorporating dilation and stride into convolution can address the limitations of long-range dependencies to some extent, it results in an unavoidable tradeoff between global and local information. On the contrary, the self-attention mechanism in Transformer layers can model the global context of images, irrespective of layer depth. A comprehensive list of transformer-based models for segmentation is provided in Table 3.

#### 8.1.1. CNN-transformer hybrids

TransUNet, proposed by Chen et al. [181], is shown in Fig. 8 and was one of the earliest examples. TransUNet uses a CNN to downsample the input image before providing it to a Transformer encoder, creating a global contextualized deep representation of the image. This representation is subsequently passed through a cascaded up-sampler to convert it into the full-resolution segmented output image. The idea of using a Transformer as a U-net encoder to learn long-range dependencies was subsequently adapted by multiple studies such as TransClaw U-Net

[188], BiTr-UNet [242], Bi-FPN-UNet [243], and Weaving Attention UNet [244]. UNet-Transformer used MHCA in skip-connections between the encoder and the decoder to recover finer spatial features [224]. LeViT-Unet [214] integrated LeViT [245] into the downsampling block of U-net. TransAttUnet [218] used a novel self-aware attention module with both Transformer self-attention and global spatial attention.

For 3D medical image segmentation, UNETR [246] used ViT-B16 [247] as the encoder instead of CNN while retaining the U-shaped network design. TransBTS used 3D CNN blocks as an encoder to model spatial information, a Transformer encoder to capture long-distance dependencies, and a decoder to model volumetric data in MRI scans [215]. CoTr concatenated CNN feature maps at different scales using positional encoding and passed them into stacked Deformable Transformer encoder blocks [248]. Deformable Transformer computed attention over a local region around reference points instead of global self-attention, reducing computational complexity. The authors showed that this methodology outperformed other CNN-Transformer hybrid models on the BCV dataset [182] that covers 11 major human organs. SpecTr [212] used adaptively sparse Transformer blocks [249] to remove redundant/noisy bands of spectral information in the Transformer encoder while segmenting hyperspectral images. This study also used 3D CNN encoders in combination with Transformer encoders in a U-Net fashion. The nnFormer [250] is a 3D Transformer for volumetric image segmentation that uses interleaved convolutional and local/global self-attention operations coupled with skip attention between the encoder and decoder to achieve better performance over other CNN-transformer hybrid models in three datasets [182, 183, 233]. Tang et al. [232] developed a new 3D Transformer-based model named Swin UNet Transformer (Swin UNETR) with a hierarchical encoder for self-supervised pre-training using five public CT datasets. The model contains a Swin Transformer encoder that directly utilizes 3D patches and is connected to a CNN-based decoder via skip connections at different resolutions. The model was fine-tuned and validated using the BCV dataset [182] and the Medical Segmentation Decathlon (MSD) dataset [233]. These studies reflect effective ways of combining convolutions with attention in medical image segmentation.

#### 8.1.2. Transformer-Only U-Nets

UTNet [191] introduced Transformer self-attention into the encoder and decoder to capture long-range dependencies at different scales. Swin-UNet [200] used pure Swin Transformer [251] blocks. The DS-TransUNet model used a dual-branch Swin Transformer in the encoder to extract feature representations at multiple scales, and Transformer Interactive Fusion (TIF) blocks to establish global interactions between them [207]. Valanarasu et al. [184] proposed Medical Transformer (MedT) with a gated axial attention layer along with local and global branches (LoGo), adapted based on position-sensitive axial attention [252] to influence positional bias on small-scale medical datasets. Karimi et al. developed a convolution-free 3D segmentation framework using pre-trained vanilla Transformer encoder, which performed better than CNN models on three proprietary datasets [253].

#### 8.1.3. Non U-Net transformer models

Zhang et al. [234] developed the TiM-Net model based on M-Net [254] with diverse attention mechanisms and weighted side output layers for retinal vessel segmentation. The model was validated on three public retinal image datasets: STARE [235], CHASEDBI [236], and DRIVE [237]. Wang et al. [238] proposed an auxiliary segmentation method for osteosarcoma detection in MRI images based on denoising and local enhancement. For noise removal, the authors used the Eformer [255]. Duc et al. [241] developed a network called ColonFormer for polyp segmentation from endoscopic images on Kvasir [194], CVC-Clinic DB [195], CVC-Colon DB [196], CVC-T [197], and ETIS-Larib Polyp DB [198] datasets. The model uses Mix Transformer [256] as the encoder backbone, a hierarchical Transformer encoder that can represent both high and low-resolution features. It also includes efficient

**Table 3**

Transformers for medical image segmentation,

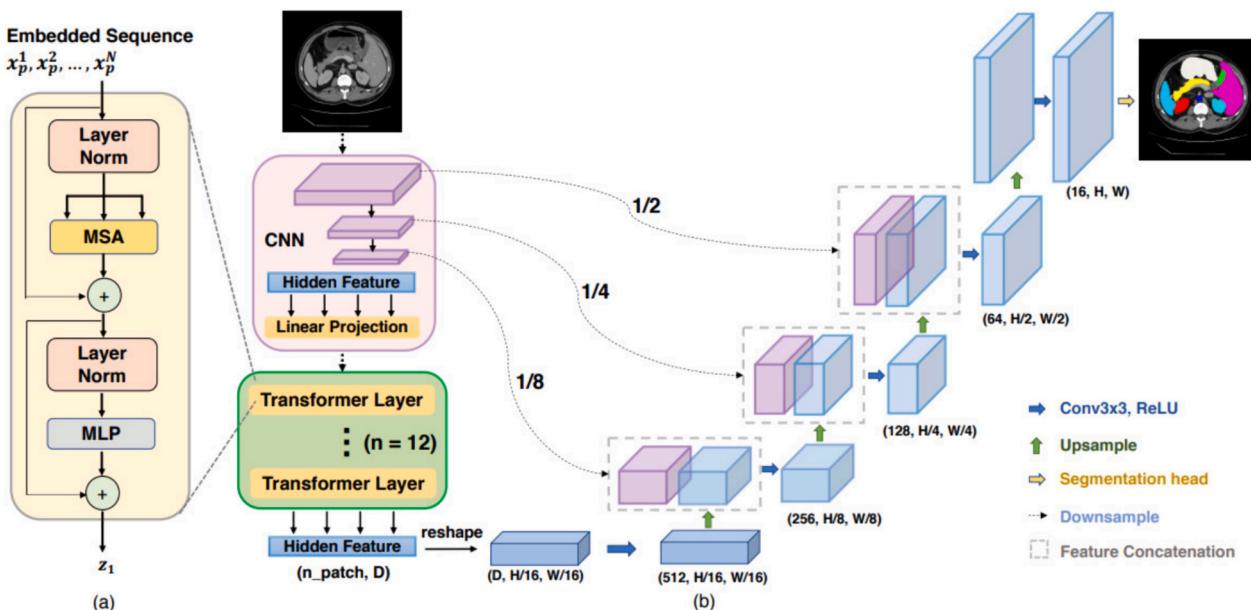
Reference	Title	Datasets	Task	Modalities
Chen et al. [181]	TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation	Synapse [182], ACDC [183]	Multi-organ segmentation, Cardiac segmentation	CT, MRI
Valanarasu et al. [184]	Medical Transformer: Gated Axial-Attention for Medical Image Segmentation	Brain Segmentation, GLAS [185], MoNuSeg [186,187]	Brain-anatomy segmentation, Gland segmentation, Nucleus segmentation	Ultrasound, Microscopy
Chang et al. [188]	TRANSCLAW U-NET: CLAW U-NET WITH TRANSFORMERS FOR MEDICAL IMAGE SEGMENTATION	Synapse [182]	Multi-organ segmentation	CT
Hatamizadeh et al. [189]	UNETR: Transformers for 3D Medical Image Segmentation	BCV [182], MSD [190]	Multi-organ segmentation	CT
Gao et al. [191]	UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation	M&Ms [192]	Cardiac segmentation	MRI
Zhang et al. [193]	TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation	Kvasir [194], CVC-Clinic [195], CVC-Colon [196], EndoScene [197], ETIS [198], BCV [182]	Polyph segmentation, Skin lesion segmentation, Hip segmentation, Prostate segmentation	Colonoscopy,
Xie et al. [199]	CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation	BCV [182]	Multi-organ segmentation	CT
Cao et al. [200]	Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation	Synapse [182], ACDC [183]	Multi-organ segmentation, Cardiac segmentation	CT
Huang et al. [201]	MISSFormer: An Effective Medical Image Segmentation Transformer	Synapse [182], ACDC [183]	Multi-organ segmentation, Cardiac segmentation	MRI
Zhang et al. [202]	Pyramid Medical Transformer for Medical Image Segmentation	GLAS [185], MoNuSeg [187], HECKTOR [203]	Gland segmentation, Nucleus segmentation	Microscopic images,
Ji et al. [204]	Multi-Compound Transformer for Accurate Biomedical Image Segmentation	Pannuke [205], CVC-Clinic [195], CVC-Colon [196], ETIS [198], Kvasir [194], ISIC2018 [206]	Tumor segmentation, Cell segmentation, Polyph segmentation, Skin lesion segmentation	CT/PET Pathology, Colonoscopy, Dermoscopy
Lin et al. [207]	DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation	CVC-Clinic [195], CVC-Colon [196], EndoScene [197], ETIS [198], GLAS [185], Kvasir [194], ISIC2018 [206]	Polyph segmentation, Skin lesion segmentation, Gland segmentation, Nucleus segmentation	Pathology, Colonoscopy, Dermoscopy
Li et al. [208]	Medical Image Segmentation Using Squeeze-and-Expansion Transformers	REFUGE2020 [209], Drishti-GS [210], RIM-ONE v3 [211], Kvasir [194]	Optic disc and cup segmentation, Polyph segmentation, Brain tumor segmentation	Colonoscopy, MRI, Fundus images
Yun et al. [212]	SpecTr: Spectral Transformer for Hyperspectral Pathology Image Segmentation	Choledoch [213]	Pathology segmentation	Pathology
Xu et al. [214]	LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation	Synapse [182], ACDC [183]	Multi-organ segmentation, Cardiac segmentation	CT
Wang et al. [215]	Transbts: Multimodal brain tumor segmentation using transformer	BraTS 2019 [216,217], BraTS 2020 [216,217]	Brain tumor segmentation	MRI
Chen et al. [218]	TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation	ISIC 2018 [206], JSRT [219], Montgomery [220], NIH [221], Clean-CC-CCI [222], GLAS [185], Bowl [223]	Chest X-ray segmentation, Skin lesion segmentation, Nucleus segmentation, Gland segmentation	X-ray, Histology, CT
Petit et al. [224]	U-net transformer: self and cross attention for medical image segmentation	TCIA, Internal dataset	Abdominal organ segmentation	CT
Yan et al. [225]	AFTer-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation	BCV [182], Thorax-85 [226], Segthor [227]	Multi-organ segmentation, Thoracic segmentation	CT
Guo et al. [228]	A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation	MSD [190]	Lung cancer segmentation	CT
Sun et al. [229]	HybridCTrm: Bridging CNN and Transformer for Multimodal Brain Image Segmentation	MRBrainS [230], iSEG-2017 [231]	Brain tissue segmentation	MRI
Tang et al. [232]	Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis	BTCV [182], MSD [233]	Multi-organ abdominal segmentation	CT

(continued on next page)

**Table 3 (continued)**

Reference	Title	Datasets	Task	Modalities
Zhang et al. [234]	TIM-Net: Transformer in M-Net for Retinal Vessel Segmentation	STARE [235], CHASEDBI [236], DRIVE [237]	Retinal vessel segmentation	Color images
Wang et al. [238]	Auxiliary Segmentation Method of Osteosarcoma in MRI Images Based on Denoising and Local Enhancement	In house dataset	Osteosarcoma segmentation	MRI
Shen et al. [239]	Dilated transformer: residual axial attention for breast ultrasound image segmentation	BUSIS [240]	Breast segmentation	Ultrasound
Thanh Duc et al. [241]	ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation	Kvasir [194], CVC-Clinic [195], CVC-Colon [196], CVC-T [197], ETIS [198]	Polyp segmentation	Colonoscopy

CT: Computed Tomography; MRI: Magnetic Resonance Imaging; PET: Positron Emission Tomography



**Fig. 8.** Overview of TransUNet architecture. a) Schematic of Transformer encoder b) TransUNet architecture. Source: The figure was adapted from [181] without modifications.

Self-Attention to reduce the computational complexity of self-attention layers.

### 8.2. Medical image registration

Image registration is the process of transforming data from multiple datasets into one coordinate system. Registration is essential for comparing, analyzing, or integrating data obtained from various sources, viewpoints, times, or sensors [257]. An example of registration is aligning CT and MRI scans of patient captured obtained from different view points and varying patient head orientation. In image registration task source and target images are provided as input to deep learning model to estimate the spatial transformation parameters between the images. Recent approaches have incorporated attention-based Transformer models for this task.

Chen et al. proposed one of the earliest Transformer based architectures, ViT-V-Net [258], to combine the vision Transformer (ViT) [247] and V-Net [259], a CNN architecture. The ViT is used to extract the features from the fixed and moving images, followed V-Net style decoder to predict the displacement field. Wang et al. [260] developed TUNet to incorporate ViT [261] into the U-Net [180] architecture to extract global and local features from moving and fixed images. Mok et al. [262] developed a fast, robust learning-based algorithm called C2FViT for 3D affine medical image registration. C2FViT leverages

global connectivity, the convolutional vision Transformer locality, and a multi-resolution strategy. Both papers evaluated their models on brain template-matching normalization and atlas-based registration using the OASIS [263] and LPBA [264] datasets. Tulder et al. proposed pixel and token-wise cross-view attention to integrate multiple views in mammography and X-ray imaging [265] using CBIS-DDSM [266] and CheXpert [267] datasets.

Chen et al. proposed TransMorph [268], a modified U-net architecture that incorporates Swin Transformer [251] blocks in its down-sampling branch for unsupervised affine and deformable image registration on the IXI [269] dataset. Transformer blocks enabled the estimation of deformation uncertainty while preserving the registration performance. Zhu et al. [270] proposed the Swin-VoxelMorph. This unsupervised learning model applies a hierarchical Swin Transformer [251] as the encoder to extract contextual information and a symmetric Swin Transformer-based decoder with a patch-expanding layer to perform up-sampling to estimate the registration fields. The authors validated the model on ADNI [271] and PPMI [272] datasets.

### 8.3. Medical image captioning and report generation

Expert medical professionals typically interpret biomedical images, documenting their findings as medical reports, a time-consuming task. Automated medical report generation can reduce the workload and

reduce human error. The image captioning/report generation tasks involve generating a textual description of a provided visual input. The input image is processed through a deep learning model to extract relevant feature information, which is fed into a language model to a coherent and contextually appropriate textual representation in the form of a sequence of words.

Hou et al. [273] proposed the RATCHET model, a medical Transformer, to generate medical text reports from chest X-rays. The authors used the MIMIC CXR v2.0.0 dataset [274] with over 300,000 chest radiograph images and free-text radiology reports. Free text reports were tokenized using the byte pair encoding approach [275]. The RATCHET architecture follows the encoder-decoder architecture, but the encoder is a CNN model, DenseNet-121 [276], whereas the decoder is the vanilla Transformer decoder. The output features of the DenseNet-121 encoder are provided as input to the second attention block of the Transformer decoder, whereby the network learns context from the radiography image against the input text report. Free text tokens are shifted right and provided as input to the decoder to predict the next token. Nicholson et al. used a pretrained ViT encoder and a pretrained PubMedBERT decoder to solve the 2021 ImageCLEFmed Caption task [277]. Their model was fine-tuned on the ROCO dataset [278] and tested on PadChest [279], CheXpert [267], ChestX-ray14 [280], and MURA [281] datasets. Alfarghaly et al. [282] used conditioned self-attention, where new key and value parameters were introduced to project the encoder's output to the decoder's attention space. The authors extracted visual and semantic features using Chexnet [283], a Densenet121 model, and pre-trained word2vec embeddings, respectively. For the training and validation of the model, they used the IU-Xray dataset [284]. You et al. [285] developed an AlignTransformer for chest X-ray images consisting of two modules: Align Hierarchical Attention (AHA) and Multi-Grained Transformer (MGT). The AHA module was used to align visual regions and disease tags. Features from the AHA module were provided as input to the MGT module. The MGT module adaptively exploited multi-grained disease-grounded visual features to determine the importance of visual features for each target word. The authors used two publicly available datasets: IU-Xray [284] and MIMIC-CXR [286]. Pahwa et al. [287] developed a memory-driven Transformer model called MedSkip for report generation. MedSkip consists of the standard Transformer encoder and a relational memory decoder. It was trained on Pathology Education Informational Resource (PEIR) Gross dataset [288] and IU X-Ray [284] datasets. Li et al. developed a Cross-modal clinical Graph Transformer (CGT) to incorporate expert knowledge into ophthalmic report generation [289]. The model first restores a sub-graph from the clinical graph and injects clinical relation triples into the visual features as prior knowledge. Reports are predicted using the encoded cross-modal features using a Transformer decoder. The CGT model was trained and validated on an ophthalmic report generation dataset called FFA-IR [290].

#### 8.4. Visual question answering (VQA)

VQA is a computer vision task where a question is posed, and the answer must be inferred from an image. In the medical domain, VQA can be used to extract information from medical images to assist in making a diagnosis. Ren & Zhou, 2020 [291] developed the CGMVQA model, which modified the original Transformer using layer normalization before the MHSA and FCN layers. The model was trained and validated on the ImageCLEF 2019 VQA-Med data set [292]. The CGMVQA can interchangeably deploy a classification or a generative mode by changing the output layer and loss function while retaining the same architecture. While in the classification mode, the model can predict a yes-no modality, in the generative mode, the model uses masked answers to predict the next word in a sentence. Naseem et al. [293] introduced the TraP-VQA model to answer medical questions presented in pathology images. This model embedded low-level visual features extracted using a CNN, low-level language features extracted using a

domain-specific language model, and the Transformer layer to learn the contextualized representation between the two to solve the VQA task. The authors used the public PathVQA dataset [294] to train and validate their model. Sharma et al., 2021 v developed an attention-based multimodal model called *MedFuseNet*, using BERT for question feature extraction, which was found to be more effective than XLNet [102] and two datasets for training: ImageCLEF 2019 MED-VQA [292] and PathVQA datasets [294].

#### 8.5. Image synthesis

Medical image synthesis aims to replace or bypass an imaging procedure constrained by time, cost, and labor or to prevent exposure to harmful ionizing radiation from some imaging modalities. It involves synthesizing medical images of a target modality from source images such as synthesizing MRI scan from CT or vice-versa. Dalmaz et al. [295] proposed a novel encoder-decoder-based generative adversarial network (GAN) model RESViT for synthesizing missing sequences in multi-contrast MRI and pelvic CT images from source MRI images. The network architecture consists of a CNN encoder, decoder, and aggregated residual Transformer to learn global representations. RESViT model synergistically fuses local and global feature representations to achieve superior image synthesis quality. Other GAN-based [296] models, such as CycleGAN [297] and CyTran [298], have been used to create contrast CT scans from non-contrast CT scans and vice versa. The CyTran architecture incorporates convolutional upsampling, convolution downsampling, and a convolution Transformer block to perform the translation. Kamran et al. [299] proposed VTGAN to combine two generators for examining local and global features with ViT [247] in a semi-supervised manner to synthesize Fluorescein Angiography images [300] while predicting retinal degeneration. VTGAN successfully synthesized angiograms from fundus images and proved robust on spatial and radial transformations.

Yan et al. created MMTrans [301] using a Swin-Transformer [251] as both a generator and registration network and a CNN as the discriminator. The generator was used to generate images with the same content as the source modality and the same style as the target modality. In contrast, the discriminator was used to measure the similarity between the original target modality images and those synthesized by MMTrans. Hu et al. proposed a double-scale graph neural network (GNN) [302] combined with a Transformer to learn long-range dependencies from global features, while for local features, they used CNN. It outperformed established baselines in the IXI dataset. Liu et al. introduced a multi-contrast multi-scale Transformer (MMT) [303], by using missing data imputation as input and proposed a Multi-contrast Shifted Window (MSwin) to capture intra- and inter-contrast dependencies.

PTNet [304], proposed by Zhang et al., synthesizes infant MRI [305] scans using a U-net [180] based architecture that incorporates a performer [306] encoder and a decoder with linear space and time complexity. PTNet outperformed previous CNN-based approaches and had an execution time of 30 slices per second. Zhang et al. further extended PTNet to 3D MRI as PTNet3D [307] and evaluated it on high-resolution Developing Human Connectome Project (dHCP) [305] and longitudinal Baby Connectome Project (BCP) datasets [308].

#### 8.6. Image reconstruction

Image reconstruction aims to reconstruct high-quality medical images with minimal cost and risk to the patient.

##### 8.6.1. Computed tomography (CT)

Low-dose computed tomography (LDCT) imaging for clinical diagnosis uses a reduced dose of X-ray radiation compared to conventional CT scans. However, LDCT is prone to noise, which affects the scan quality. Zhang et al. proposed TransCT [309] to enhance the quality of LDCT images using the AAPM-Mayo LDCT dataset [310]. The input

image was decomposed into low-frequency and high-frequency components, and then the content, texture, and high-frequency embeddings were fed to the TransCT model to obtain refined high-frequency textural features. Luthra et al. proposed Eformer [255] by combining learnable edge-enhancement convolutions called Sobel filters and the LeWin transformer [311] in denoising LDCT images for detecting metastatic liver lesions (AAPM-Mayo dataset) [310]. Wang et al. [312,313] proposed convolution-free transformer-based encoder-decoder dilation networks (TED-net) using vanilla transformer blocks for LDCT denoising. Instead of an image, a few approaches used informative sinograms generated by restoration modules from origin LDCT images for reconstruction using Transformer-based models [314–317].

#### 8.6.2. Magnetic resonance imaging (MRI)

Korkmaz et al. proposed an MRI reconstruction model based on a zero-shot learned adversarial vision Transformer named SLATER [318] to overcome the data size limitation. Inspired by Deep Image Prior (DIP) [319], they replaced the CNN backbone of DIP with a cross-attention Transformer and outperformed DIP on the IXI dataset [269] and fastMRI dataset [320]. Feng et al. [321,322] introduced a multi-task framework T2Net, to share the representations between reconstruction and super-resolution branches. Furthermore, they extended to multi-modalities (MTrans), aiming to learn more knowledge from MRI using both branches. Fang et al. proposed a cross-modality high-frequency Transformer (Cohf-T) [323] for super-resolving, low-resolution MR images. Guo et al. proposed a lightweight recurrent Transformer model ReconFormer [324], which includes pyramid transformer layers [325] to capture intrinsic multiscale information and feature correlation through the recurrent states. Li et al. proposed McMRSR [326], a Transformer based network to model long-range dependencies between reference and target images and aggregate multiscale matched features to reconstruct a target MR image. Few approaches use raw K-space signals of MRI scans instead of final MRI images as they contain learnable information for MRI reconstruction [320,327–330]. Hu et al. introduced a Transformer-enhanced Residual-error AlterNative Suppression Network [331], which included a regularization term to improve the contribution of high-frequency information during inference. Fabian et al. [332] proposed HUMUS-Net, a two-level hybrid CNN Transformer architecture for MRI reconstruction using the fastMRI dataset [320]. Huang et al. [333] proposed a GAN [296] based on Swin-Transformer [251] named ST-GAN, which preserved edge and texture features. Swin-Transformer inspired shifted window attention became the go to Transformer architecture for many studies targeting MRI reconstruction [328,334–336].

#### 8.6.3. Positron emission tomography (PET)

PET is an imaging technique that measures emissions from radioactively labeled chemicals injected into the bloodstream. PET scans can measure metabolic activity and other biochemical functions. Unfortunately, PET suffers from a poor signal-to-noise ratio, and its reconstruction requires denoising low-quality PET images to create high-quality ones. Luo et al. proposed a GAN based Transformer model, Transformer-GAN [337], for PET reconstruction with CNN(Encoder)-Transformer-CNN(Decoder) architecture to take advantage of spatial information and long-range dependencies from CNN and transformers, respectively. Fu et al. extended their transGAN-SDAM [338] for fast 2.5D-based L-PET. The transGAN generates higher quality F-PET images, followed by the SDAM module, which combines spatial information of an F-PET slice sequence to generate whole-brain F-PET images. Jang et al. proposed Spach Transformer [339] that can leverage spatial and channel-wise information based on local and global MHSA, which outperformed baselines on different PET tracer datasets of 18F-FDG, 18F-ACBC, 18F-DCFPyL, and 68GaDOTATATE.

## 9. Transformers for critical care

### 9.1. Predicting long-term adverse outcomes

Transformers have been used to predict adverse outcomes after critical care such as recurrence or death. Yang et al., 2021 predicted a 60-day and 90-day response to targeted immunotherapy of patients with non-small cell lung cancer (NSCLC) using asynchronous clinical time series consisting of chest CT scans, and blood tests, and patient characteristics using an attention module called Simple Temporal Attention [340]. The model predicted which patients would have long-term durable survival gains under an immunotherapy regimen. Similarly, in 2021 for colorectal cancer, Ho et al. used Transformer encoders to extract features from sequential carcinoembryogenic antigen (CEA) measurements. It combined CEA measurement features with deep representations of tabular features such as tumor sites, number, dates, and dosage of chemotherapy to predict recurrence [341]. They modified the Transformer to incorporate 1D convolutions prior to localized self-attention [342]. Their model outperformed commercial diagnostic tests of colorectal cancer recurrence. Non-clinical population-level claims data has also been modeled using multi-headed self-attention to predict relapse after surgery [343,344]. These studies utilized the French national health insurance database (SNIIRAM), consisting of health-insurance claims entries of 65 million individuals [345].

### 9.2. Surgical instruction generation

Intra-operative surgical assistance AI systems need to solve the task of automatic surgical instruction generation. Zhang et al., 2021 used a Transformer-backboned encoder-decoder network combined with self-critical reinforcement learning (RL) to jointly model surgical activity and relationships between visual information and textual description [346]. They used the Database for AI Surgical Instruction dataset (DAISI) to evaluate their model [347]. The authors used a combination of machine translation and image-captioning criteria to evaluate their models, such as BLEU [348], Rouge-L [349], METEOR [350], and CIDEr [351], and SPICE [352]. The combination of Transformer with RL beat baselines comprising LSTM-based fully connected and soft-attention models.

## 10. Transformers for social media data in public health

In recent years, using social media data has gained prominence in different areas of public health [353–356]. Transformers have been applied to social media data for addressing several public health problems, such as monitoring adverse drug reactions [357,358], monitoring mental health [359], categorizing vaccine confidence [360], and locating disease hotspots [361]. In this section, we present the models and their performance on social media datasets.

### 10.1. Monitoring adverse drug events (ADEs)

ADEs, refers to an undesired, unpleasant, or dangerous reaction to a medication [362], which has been found to be underreported; thus, researchers have recently used social media to improve ADE monitoring [363]. The main steps in monitoring ADRs using social media posts are text classification to find ADE mentions, followed by extracting the ADE concept and mention from the classified text. Breden et al. [357], pre-processed the Twitter dataset from Social Media Mining for Health (SMM4H) 2019 Competition [364] using the lexical normalization [365] method. The best-performing model was an ensemble of fine-tuned BERT, BioBERT [54] and ClinicalBERT [59]. Sakhovskiy et al. [366] used a more recent dataset provided by SMM4H 2021 [367] for task1, classifying English tweets by concatenating the RoBERTa [95] and ChemBERTa [368] models. For task2, Russian tweets classification performed by concatenating the EnRuDR-BERT [369], and

ChemRoBERTA [362] cross-attention. Hussain et al. [370] proposed an end-to-end system based on transfer learning using one prediction head for the text classification and another for labeling the adverse drug responses. The authors fine-tuned BERT with a modular Framework for Adapting Representation Models (FARM) and present the FARM-BERT framework, which outperforms competing models on TwiMed-Twitter [371], Twitter [372], PubMed [373], and TwiMed-PubMed [371] datasets. The framework FARM-BERT supports multitask learning by combining multiple prediction heads, making training of the end-to-end systems easier and computationally faster. Raval et al. [358], tackled the same ADE classification problem; however, they framed it as a sequence-to-sequence problem and used the pre-trained T5 model architecture [374] on multiple datasets (SMM4H [375], CADEC [376], ADE corpus v2 [373], WEB-RADT [377], SMM4H-French [375]). The authors further expanded the proportional mixing and temperature scaling training strategies described in [378] to handle multi-dataset and present relative improvement on the F-1 score.

#### 10.2. Monitoring depression

A large-scale depression dataset on Twitter was presented by Zhang et al. [359], used Transformer-based models to identify users suffering from depression using their everyday speech. The importance of psychological text features was also studied when performing depression classification. Results on the fluctuating depression levels for different groups were also presented. Matero et al. [379] used pretrained BERT embeddings to encode this information. Kabir et al. [380] presented a dataset observing the severity of depression in tweets and reported baseline results using BERT and DistilBERT [381].

#### 10.3. Monitoring diabetes

Large-scale Twitter data concerning diabetes-related tweets have been collected and used to identify cause-effect relationships [382]. They used a pre-trained BERTweet model [383] to detect causal sentences and a combined BERT+ Random Field Generator model to extract potential cause-effect relationships.

#### 10.4. Categorizing vaccine confidence

Social media plays a crucial role in gauging public discourse on topics such as vaccine effectiveness [384]. It provides a proxy to analyze vaccination apprehensions and study the barriers to successful vaccinations [385]. Kummervold et al. [360] used domain-specific BERT model to assess the social media stance towards vaccination during pregnancy on a dataset of 2722 unique tweets. The model was able to achieve accuracy of a trained human annotator in categorizing the stance, outperform other models and human coders in some cases.

#### 10.5. Locating disease hotspot

It is essential to detect disease outbreaks while simultaneously reducing reporting lag time. This can provide another source of data to complement traditional surveillance approaches. Alsudias et al. [361] performed a multi-label classification task to identify tweets of infected individuals in the Arabic-speaking world. The authors propose a combination of binary relevance, classifier chains, label power set, multi-label adapted k-nearest neighbors (MLKNN) [386], support vector machine with naive Bayes features (NBSVM) [387], BERT and AraBERT (transformer-based model for Arabic language understanding) [388]. The proposed model achieved an F1 score of up to 88% in the influenza case study and 94% in the COVID-19. It is shown that including informal terms and non-standard terminology (e.g., the slang term of influenza, symptom, prevention, treatment, infected with) in the encodings improved the performance by as much as 15%, with an average improvement of 8%. The proposed geolocation detection algorithm

performed moderately in predicting the location of users according to their tweet content.

### 11. Monitoring bio-physical signals

Transformers have been used to model physical activity, Electroencephalogram (EEG), and Electrocardiogram (ECG) signals. In the following sections, we review these works.

#### 11.1. Human activity recognition (HAR)

Human Activity Recognition (HAR) is a proliferating field of research owing to the recent rise of wearables, smartphones, and Internet of Things (IoT) devices. Some studies have used multimodal self-attention to fuse features from various modalities [389,390]. They studied sequences of human movements through multimodal data (such as RGB, depth, and skeletal data) [391–393] or modeled human activity through accelerometers and gyroscopes [394–397]. Spatiotemporal bone and joint sequences from skeleton data have been modeled using multi-scale Transformers on multiple datasets [398–401]. Owing to the lack of simple augmentation strategies of longitudinal sensor data, Ramachandra et al. used Transformer-GAN to provide a speedup over existing Recurrent-GAN [402].

#### 11.2. Electroencephalogram (EEG)

Electroencephalogram (EEG) is a widely used noninvasive measurement of brain activity. Transformers have been used to classify visual or motor imagery using EEG signals [403]. It has been shown that extensive self-supervised pre-training using contrastive loss can help Transformer models represent EEG data collected using different hardware while performing different tasks [404]. Pretraining was conducted using the Temple University Hospital EEG Corpus [405] and downstream analyses were done using a battery of smaller datasets [406–408]. Cross-modal Transformers have been used to find contextualized embeddings representing associations between auditory attention detection and EEG signals [409]. This can disentangle sources of brain activity at different time points while the subject is attending to multiple sound sources simultaneously. This study was conducted on the Denmark Technical University (DTU) dataset [410,411]. Finally, a 2D Transformer was used to capture local self-similarity, and feed-forward connections were used to capture global self-similarity to create a novel denoising system for 1D EEG signal [412] using another publicly available dataset [413].

#### 11.3. Electrocardiogram (ECG)

Electrocardiogram (ECG) signals alone and combined with other sensory information were used to predict stress in subjects using Transformers [414,415]. The Wearable Stress and Affect Detection (WESAD) and SWELL Knowledge Work (SWELL-KW) are publicly available datasets used for this purpose [416,417]. A transformer network embedded inside a CNN architecture has been used to classify arrhythmia [418].

### 12. Transformers for biomolecular sequences

Biomolecular sequences can represent genomic, proteomic, and drug data. As sequence translation models, transformers have been widely used to model the relationships between anomalous biological sequences and related diseases. Moreover, drug/protein synthesis or gene sequence alignment problems have been treated through the lens of machine translation, where the Transformer is the model of choice.

### 12.1. DNA

Gene Transformer, which consists of a multi-head self-attention module, detects lung cancer subtype biomarkers [419]. It consists of two 1D convolutional layers before the MHSA layer to extract low and moderate-level features. A previous study utilized RNA-sequencing values from lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) datasets from the Cancer Genome Atlas project [420]. Clauwaert et al. 2020 introduced an attention method optimized for nucleotides on top of the Transformer-XL architecture [421]. This attention module included a 1D convolutional layer that extracted overlapping DNA segments of length k called k-mers from the original DNA sequences' query, key, and value matrices. The authors solved three problems, including a) annotating the transcription start site (TSS), b) annotating the translation initiation site (TIS), and c) recognizing 4mC methylation sites using the following datasets – RegulonDB [422], Ensembl [423], and MethSMRT [424], respectively. A following study utilized comparative TSS annotations from multiple datasets, including RegulonDB [422], Etwiller, et al., 2016 (Cappable-seq) [425], Yan et al., 2018 (SMRT-Cappable-seq) [426], and Ju et al., 2019 (SEnd-seq) [427]. In another study, the Transformer-XL network was highly biased toward attending to promoter regions and transcription factor binding sites near the gene under question [428]. Another network, DNABERT was used to predict transcription factor binding (TFB) sites, including proximal and core promoter regions, splice sites, and genetic variants [429]. Reference human genome GRCh38.p13 primary assembly from GENCODE Release 33 [430] was used for pre-training, TATA, and non-TATA promoter data from Eukaryotic Promoter Database (EPDnew) [431] for promoter prediction and ENCODE 690 ChIP-seq profiles from UCSC genome browser [432] were used for predicting TFB sites. Enhancers are regulatory elements that activate promoter transcription over large distances independently of orientation [433]. BERT, pre-trained with masked language modeling (MLM) and the next sentence prediction tasks, was combined with 2D convolutions to predict transcription enhancers using a dataset describing an enhancer sequencer from nine cell lines [434,435].

### 12.2. Protein

Transformers can either predict global properties of protein such as type, function, or cellular localization or infer local properties of selected protein residues such as 2D/3D structure or post-translation modifications (such as phosphorylation and cleavage sites) [436]. The recent success of AlphaFold in protein structure prediction problems [437] has significantly changed the domain [438], although recent advances have primarily included fine-tuning pre-trained deep models for learning with small datasets [436].

### 12.3. Molecular drugs

Transformer have been utilized for the prediction of molecular drugs as follows.

#### 12.3.1. Drug-drug synergy

One of the most useful applications of Transformer networks is in the finding of synergistic combinations of drugs for the treatment of diseases which cannot be cured by a single molecule. The classic example of this is cancer. In cancer, drug combinations alleviate drug resistance and improve therapeutic efficacy. However, the rapidly growing number of anti-cancer drugs makes it extremely resource intensive to search the entire space of synergistic drug combinations. This is where computational models like the Transformer are useful. The TranSynergy model constructed a Transformer model of the cellular effect of drug combinations on different gene-cell line combinations by modeling cell-line gene dependency, gene-gene interaction, and genome-wide drug-target interaction, thereby introducing mechanistic knowledge into the

model [439]. The study utilized a large drug synergy score dataset [440] and drug target profiles from DrugBank[441] and ChEMBL[442]. TranSynergy outperformed the SOTA and predicted multiple novel synergistic drug combinations for treating ovarian cancer. Kim et.al., 2020 used multi-task transfer learning to study drug synergy in under-studied tissues to overcome data scarcity problems [443]. The authors used a multi-head Transformer network to create an embedding of the Simplified Molecular-Input Line-Entry System (SMILES) representation of drugs. TP-DDI presents a completely end-to-end Transformer pipeline with pretrained BioBERT weights for drug recognition and drug-drug interaction (DDI) classification [444]. This study is conducted on the DDI Extraction 2013 corpus [86] which consists of a list of semantically annotated documents with sentences referring to drugs and DDIs from the DrugBank database and MedLine abstracts.

#### 12.3.2. Drug synthesis

Transformers have been used to convert the task of target-driven de novo drug-synthesis into a neural machine translation task that converts an amino acid sequence into the chemical formula of its binding drug [445]. This method needs neither any prior information about the drug structure nor the 3D structural information of the protein target. The study used a dataset of binding affinity between proteins and drug-like molecules from the BindingDB database [446]. Synthesized drugs were evaluated on active properties like the number of hydrogen donors/acceptors, molecular weight, length, total polar surface area, number of rotatable bonds, and drug-likeness. Born et.al., 2021 studied the synthesis feasibility of drugs for use against the SARS-CoV-2 virus using a transformer-based retrosynthesis prediction engine [447] consisting of two molecular transformers [448]. They operate on a SMILES representation of a molecule to predict best routes for its synthesis [449]. This information was further utilized by another Transformer model to predict the optimal synthesis protocol using a text representation of the synthesis steps [450]. The approach incorporated variational autoencoders and reinforcement learning to automatically learn molecules that target ACE2, a surface receptor on human epithelial cells that allows entry of the SARS-CoV-2 virus [449].

#### 12.3.3. Drug-target interactions

In-silico drug discovery is driven by computational models of drug-target interactions. Huang et al. developed the Molecular Interaction Transformer, which models the interaction space between the most common substructures of molecules and drugs [451]. These substructures were discerned using Frequent Consecutive Sub-sequence algorithm on protein sequences from UniProt dataset [452] and drug SMILES strings from ChEMBL [453]. In this work, a Transformer encoder is used to create contextualized embeddings of protein and drug substructures separately which are multiplied to capture their interaction strengths. A CNN extracts higher order interactions from joint space. Three datasets were employed to learn the transformer and CNN weights- MINER DTI from BIOSNAP [454], BindingDB [455] and DAVIS [456].

Manica et al., 2021 proposed an anticancer drug sensitivity model using drug SMILES sequences, gene expression profile of tumors, and protein-protein interaction networks.[457] In this model, an attention-based gene expression encoder generates self-attention weights, a contextual attention layer ingests this gene embedding together with the SMILES encoding of a drug to compute an attention distribution over the SMILES tokens, in the genetic context. CNNs with variable kernel lengths were used to extract information about all possible substructures inside the SMILES sequence. The model outperformed others on a regression task involving prediction of drug IC50 values. Training was done using lenient splitting which prevented cell-drug pairs in the test data from being seen beforehand but did not prevent the model from observing how a given cell interacted with other drugs in the dataset and vice versa. The authors used drug sensitivity data from the publicly available Genomics of Drug Sensitivity in Cancer (GDSC) database for

this study [458].

Morris et.al. 2020 proposed a transformer-based machine translation method to inform the segmentation of molecular substructures into binding/non-binding a target protein [459]. The authors translated SMILES encodings to IUPAC nomenclatures for a set of 83 million compounds from PubChem [460] database and used the resultant cross-representation attention embeddings as features to classify binding/non-binding compartments of molecules from BindingDB [446] to important proteins including HIV-1 protease.

#### 12.3.4. Drug metabolism prediction

Metabolic processes in the human body can change a drug's structure, diminishing its safety and efficacy. Therefore, investigation of the metabolic effect of a candidate drug is crucial in drug design studies. Litsa et al., 2020 fine-tuned a pretrained Molecular Transformer, and used an ensemble of them with beam search to find k-likeliest metabolites from every drug [461]. The Molecular Transformer [448] was pretrained on this dataset [462] consisting of 900,000 training instances. The network was further fine-tuned using a manually curated dataset combining samples from Drug-Bank (version 5.1.5) [441], Human Metabolome Database (HMDB) (version 4.0) [463], HumanCyc from MetaCyc (version 23.0) [464], Recon3D (version 3.01) [465], the biotransformation database (MetXBioDB) [466] and reaction rules from SyGMA [467]. Their network outperformed SOTA models including the BioTransformer [466].

### 13. Discussion

This paper presented an exhaustive summary of Transformer-based applications in healthcare for tasks such as clinical report generation, medical image segmentation and registration, molecular sequencing, drug-drug interactions, protein synthesis, surgical augmentation, and bio-physical signal analysis. Although relatively new, Transformers have become remarkably successful due to several reasons such as, parallelizable attention computation, ability to model long range dependencies, scalability, transfer learning, ability to produce contextual embeddings, interpretability and universal adaptability to various data modalities beyond text data. However, the parallelizable attention module at the heart of the Transformer network is computationally expensive and often needs to be optimized for efficient usage. In what follows, we highlight potential drawbacks of transformers, how to overcome them, and new directions enabled by Transformers.

#### 13.1. Interpretability and explainability

Most deep learning systems are considered “black box” models because their inferences do not come with any discernable explanation. This lack of interpretability has traditionally prevented the systemic acceptance of AI-aided diagnostics in the medical domain. Transformers inherently provide some transparency through visualization of their attention weights. Trained attention weights elucidate contextual information significant for downstream inference. However, interpreting Transformers is challenging due to the frequent use of skip-connections and the dynamic nature of the model, which involves weight computation through matrix multiplication. Therefore, Transformer interpretability, albeit being an inherent property, is not trivial. Chefer et al. [468] show that Transformer attention is often fragmented and does not provide a robust explanation. They also proposed a novel way to compute relevancy based deep Taylor decomposition principle and propagate the scores through the transformer layers. In case of vision Transformers, Bohle et al. [469] proposed B-cos transformers, for holistic explanations for their decisions while retaining the performance to the baseline ViTs. Disease diagnosis prediction studies [470,471] have generated attention visualizations and cosine similarity between the learnt clinical diagnoses embeddings verified by expert clinicians to understand whether the trained model could capture the underlying

semantic of diagnoses codes. However, there remains a need to develop novel techniques to improve the interpretability of Transformer models tailored towards healthcare AI.

#### 13.2. Environmental impact

Advances in AI in recent years have come at the cost of a massive carbon footprint. Training a large-scale deep learning model is estimated to produce 626,000 lbs of carbon dioxide, equivalent to five automobiles' lifetime emissions [472]. The number of computational resources researchers use to create SOTA models has doubled every three to four months [473]. Most emissions are associated with developing and training deep learning algorithms, whereas fine-tuning and adaptation contribute less [474]. Strubell et al. [472] suggested that researchers report hardware-independent training time measurements, such as the number of gigaflops required for training convergence and measuring model sensitivity to data and hyperparameters. The last decade has seen advancements in AI-augmented healthcare, on the one hand, and carbon emissions caused by AI systems that are detrimental to the climate and public health on the other. Large healthcare conglomerates and governmental agencies around the world should target net-zero carbon emissions. United Kingdom National Health Service has set a goal of net-zero emissions by 2040 [475]. Goals such as this are vital to promote the development of energy-efficient hardware and algorithms that make AI sustainable and globally accessible.

#### 13.3. Computational costs

The reason behind the impact of Transformers is their high parametric complexity, flexibility to handle unequal input lengths and model scalability. However, Transformers' ability to be trained on enormous datasets comes with expensive computational training budgets. The LLM GPT-3 [23] by OpenAI training is estimated to cost \$4.6 million and 355 years of computing time using the Nvidia Tesla V100 device [476]. Google's 530 billion parameters PaLM model is estimated to consume 103,500 KWh over 60 days [477]. Training and deploying large-scale AI models with high-end hardware requirements in healthcare settings is challenging. For example, for on-premise use in a hospital, a centralized compute cluster similar to ChatGPT might need to be maintained and interacted with using an API. However, healthcare settings typically need lightweight models to generate real-time predictions with minimal maintenance costs. Techniques for compressing deep learning models, such as pruning [478], knowledge distillation [479], and quantization [480], can be used to provide a more efficient model implementation for deployment within practical hardware constraints.

##### 13.3.1. Model compression

Transformer models can be efficiently compressed by discarding some attention heads during the inference phase. Michel et al. [481] showed that models trained on multiple heads during training time need not require all the heads during test time. Similar redundancy has been observed in generating attention matrices from multiple heads [482].

##### 13.3.2. Quantization

Quantization-based approaches reduce the number of bits/unique values required to represent model weights and intermediate layer activations. There has been growing interest among researchers in recent years in quantizing transformer networks. Shen et al. [483] observed ~2.3% degradation in performance with quantization down to 2 bits, corresponding to 13X compression of network parameters and 4X compression on embeddings and activations. It was observed that position embedding and the embedding layers are more sensitive to quantization than other operations.

##### 13.3.3. Knowledge distillation

The knowledge distillation approach aims to train small networks

(aka student) using the knowledge from the large model (teacher). Student models are obtained by reducing encoder width, number of heads, and number of encoders and replacing them with CNN, BiLSTM, or a combination [484]. Dimensional incompatibility between the student and teacher due to compact representations can be overcome by projecting teacher or student outputs [485]. Sun et al. [479] proposed patient knowledge distillation to compress large teacher BERT model trained on MIMIC-III dataset into shallow student models. Student models patiently learned from intermediate layers, which translated into improve performance and significant training-efficiency gain.

#### 13.3.4. State space models

Transformer self-attention is capable of handling intricate interactions among sequence elements. However, this capability presents a limitation when applied to exceedingly long sequences, particularly in modalities like audio, video, and accelerometry where data extends continuously over time. State space sequence models [486], on the other hand, state space models excel in modeling long range sequences while maintaining computational efficiency. Conceptually, state space models can be seen as a fusion of recurrent neural networks and convolutional neural networks, offering linear or near-linear scalability to sequence length. A recent state space model named Mamba [487] has introduced a selection mechanism within its architecture, allowing it to make informed decisions about the information to propagate or discard based on its relevance to tokens in the sequence. Mamba leverages a hardware-efficient implementation inspired by FlashAttention [488], resulting in a remarkable 5X faster inference speed compared to Transformers. Mamba outperformed transformers of same size and matched the performance of transformers twice its size.

#### 13.4. Fairness and bias

A model is biased when it exhibits undesired dependence on an attribute of the data that belongs to a specific demographic group [489], and could lead to unfair treatment of particular patient groups. Researchers have observed that bias often arises when the datasets used to train the models under-represent certain patient populations [490–492]. Although this is a prevalent bias problem during training, other sources of bias at all stages exist, including during problem formulation, data collection, data preprocessing, model development and validation, and model deployment (e.g., due to unmonitored drift) [493]. With the increasing scale of models and amount of data available, the existing biases and stereotypes perpetuate into the models leading to unfair and biased outcomes [49]. Thorough validation should be done before deploying the model to evaluate the performance of underrepresented groups. The models should be continuously monitored and audited for fairness and bias post-deployment.

#### 13.5. AI alignment

The goal of AI alignment is even broader than preventing bias by striving to design AI systems that align with human values and goals. An AI system is considered aligned when the system behaves in ways beneficial to humans while minimizing the risk of unintended consequences and harmful outcomes [36]. LLMs sometimes confidently assert false claims that do not reflect facts, a phenomenon termed hallucination [494]. These hallucinations by the misaligned models fail to meet the user's expectations of correct answers faithful to the existing sources. Ensuring AI systems are aligned with human values and goals is challenging because predicting and designing for every potential desired and undesired outcomes is difficult. As AI systems become more capable, they become increasingly susceptible to the alignment problem, which can result in unintended and harmful consequences [495]. AI alignment is especially critical in healthcare when deploying large-scale foundation models to ensure these models are ethical, responsible, respectful of patient privacy, and, most importantly, not causing harm. Healthcare

professionals and the AI research community need to develop a clear set of standards and guidelines to establish ethical use of AI in health care.

#### 13.6. Data privacy and data sharing

Preserving patient privacy is a required feature in all healthcare AI systems. Federal regulations based on the Health Insurance Portability and Accountability Act (HIPAA) regulate the development of AI models that use patient information [496,497]. Nonetheless, this also adversely impacts the development of large models such as Transformers that require large amounts of data. Utilizing data from a few sources, such as select public repositories, can skew the model inferences based on underlying limitations in dataset collection (different equipment, protocol, and cohort demographics), processing (specific heuristic or statistical preprocessing), and deployment (different metadata, availability, and maintenance). These biases can skew predictions that favor or adversely affect certain population groups over others, leading to a degradation in the quality and equity of healthcare for individuals from the protected group and stymieing the research on age, sex, or race-related medical conditions.

The Federated learning (FL) paradigm shown in Fig. 9 aims at developing a shared training model that can leverage data from multiple fragmented sources, such as different healthcare institutions, without divulging sensitive patient information [498]. FL communicates between various data sources by exchanging model-specific characteristics like parameters and gradients without exchanging patient information directly. Recent efforts in FL have targeted digital health objectives like determining patient clinical similarity [499,500], mortality and ICU length-of-stay [501], brain segmentation [502], and brain-tumor segmentation [503,504]. FL can perpetuate many healthcare innovations in the future. However, there are technical challenges in building an operational FL workflow, such as inhomogeneous data distributions, computational hardware differences, inconsistent privacy preservation settings, and resultant performance trade-offs [505].

### 14. Conclusion

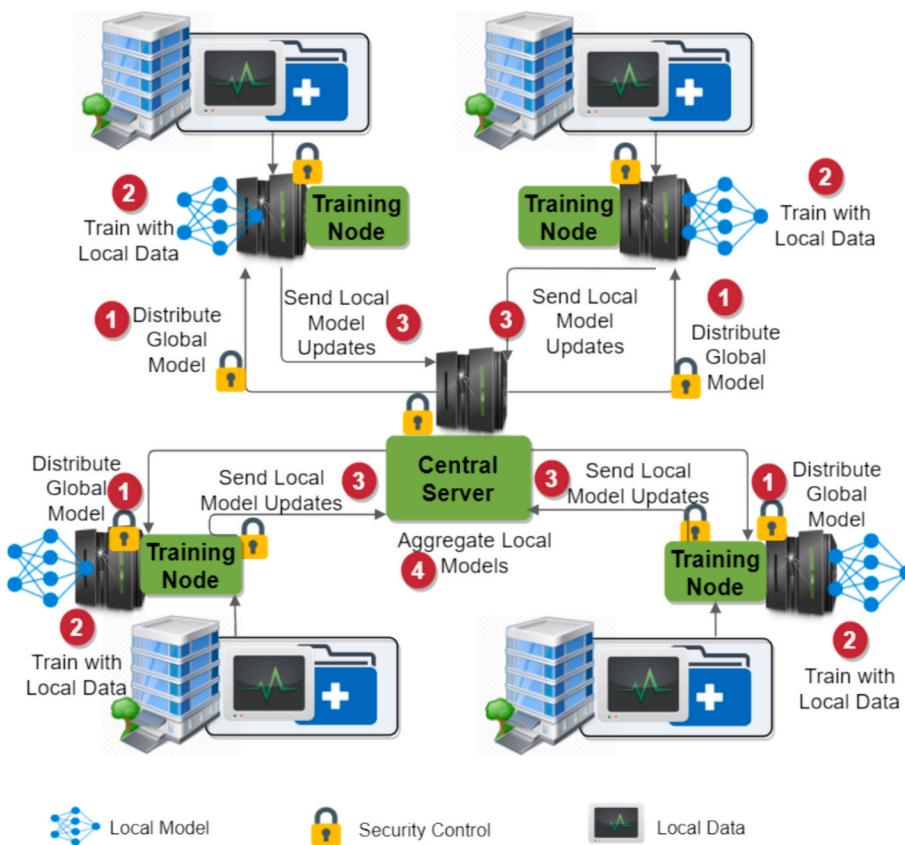
Transformer models have demonstrated enormous potential in a wide variety of healthcare applications. They possess a unique ability to model various data modalities, including images, clinical text, biophysical signals, structured EHR, social media and genomic data. From disease diagnosis to drug discovery, Transformer models exhibit the potential to improve patient outcomes and advance medical research. However, various challenges and limitations remain to be addressed before they are widely accepted into regular clinical practice. These include data limitations, biases, privacy, security, and truthfulness. The majority of the models currently in use are task-specific, and there is a need to utilize robust multimodal inputs in many cases. Nevertheless, the future of AI in healthcare is optimistic, with promising advancements and opportunities presented by large-scale transformer models.

### Funding

This study is supported by National Science Foundation CAREER award 1750192, 1R01EB029699 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), and 1R01NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS).

### CRediT authorship contribution statement

**Subhash Nerella:** Writing – original draft, Visualization, Conceptualization. **Sabyasachi Bandyopadhyay:** Writing – review & editing, Writing – original draft, Visualization. **Jiaqing Zhang:** Writing – original draft, Visualization. **Miguel Contreras:** Writing – original draft.



**Fig. 9.** Schematic of Federated learning with a central server that interacts with training nodes at different locations continuously updating the model parameters without exchanging the data between local and central servers.

**Scott Siegel:** Writing – original draft. **Aysegul Bumin:** Writing – original draft. **Brandon Silva:** Writing – original draft. **Jessica Sena:** Writing – original draft. **Benjamin Shickel:** Writing – review & editing. **Azra Bihorac:** Writing – review & editing, Funding acquisition, Conceptualization. **Kia Khezeli:** Writing – review & editing. **Parisa Rashidi:** Writing – review & editing, Supervision, Conceptualization.

#### Declaration of competing interest

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

#### References

- [1] "The healthcare data explosion, [https://www.rbccm.com/en/gib/healthcare/episode/the\\_healthcare\\_data\\_explosion](https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion)(accessed Feb. 5, 2022).
- [2] Vaswani A, et al. Attention is all you need. *Adv Neural Inf Proces Syst* 2017; 5998–6008.
- [3] Wolf T, et al. "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020. p. 38–45.
- [4] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *nature* 1986;323(6088):533–6.
- [5] Khan S, Naseer M, Hayat M, Zamil SW, Khan FS, Shah M. "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*. 2021.
- [6] Liu Y, et al. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*. 2023.
- [7] Aleissae AA, et al. Transformers in remote sensing: A survey. *Remote Sens* 2023; 15(7):1860.
- [8] Q. Wen et al. "Transformers in time series: A survey," arXiv preprint. [arXiv:2202.07125](https://arxiv.org/abs/2202.07125); 2022.
- [9] Latif S, Zaidi A, Cuayahuitl H, Shamshad F, Shoukat M, Qadir J. "Transformers in speech processing: A survey," arXiv preprint. [arXiv:2303.11607](https://arxiv.org/abs/2303.11607); 2023.
- [10] Xu P, Zhu X, Clifton D. Multimodal Learning With Transformers: A Survey" in *IEEE Transactions on Pattern Analysis & Machine Intelligence*45(10); 2023. p. 12113–32.
- [11] Shamshad F, et al., "Transformers in Medical Imaging: A Survey," arXiv preprint. [arXiv:2201.09873](https://arxiv.org/abs/2201.09873); 2022.
- [12] He K, et al., "Transformers in medical image analysis: A review," arXiv preprint. [arXiv:2202.12165](https://arxiv.org/abs/2202.12165); 2022.
- [13] Parvait A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision Transformers in medical computer vision—A contemplative retrospection. *Eng Appl Artif Intell* 2023;122:106126.
- [14] Wang B, Xie Q, Pei J, Tiwari P, Li Z. "Pre-trained language models in biomedical domain: A systematic survey," arXiv preprint. [arXiv:2110.05006](https://arxiv.org/abs/2110.05006); 2021.
- [15] Harzing AW. "Harzing, A.W. (2007) Publish or Perish," ed. 2007.
- [16] Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia, ed 2023.
- [17] He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
- [18] Lin T, Wang Y, Liu X, Qiu X. "A survey of transformers," arXiv preprint. [arXiv:2010.04554](https://arxiv.org/abs/2010.04554); 2021.
- [19] Tay Y, Dehghani M, Bahri D, Metzler D. "Efficient transformers: A survey," arXiv preprint. [arXiv:2009.06732](https://arxiv.org/abs/2009.06732); 2020.
- [20] Devlin J, Chang M-W, Lee K, Toutanova K. "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805); 2018.
- [21] Bommasani R, et al., "On the opportunities and risks of foundation models," arXiv preprint. [arXiv:2108.07258](https://arxiv.org/abs/2108.07258); 2021.
- [22] Chowdhery A, et al., "Palm: Scalable language modeling with pathways," arXiv preprint. [arXiv:2204.02311](https://arxiv.org/abs/2204.02311); 2022.
- [23] Brown T, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst* 2020;33:1877–901.
- [24] Touvron H, et al., "Llama: Open and efficient foundation language models," arXiv preprint. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971); 2023.
- [25] Anil R, et al., "Palm 2 technical report," arXiv preprint. [arXiv:2305.10403](https://arxiv.org/abs/2305.10403); 2023.
- [26] OpenAI R. "GPT-4 technical report," arXiv, p. 2303.08774. 2023.
- [27] Taylor R, et al., "Galactica: A large language model for science," arXiv preprint. [arXiv:2211.09085](https://arxiv.org/abs/2211.09085); 2022.
- [28] Wu S, et al., "Bloomberggpt: A large language model for finance," arXiv preprint. [arXiv:2303.17564](https://arxiv.org/abs/2303.17564); 2023.
- [29] "Openai codex. [https://openai.com/blog/openai-codex.](https://openai.com/blog/openai-codex/)," ed (accessed Jan. 11, 2023).
- [30] Bowman SR. "Eight things to know about large language models," arXiv preprint. [arXiv:2304.00612](https://arxiv.org/abs/2304.00612); 2023.

- [31] Wei J. et al., "Emergent abilities of large language models," arXiv preprint. [arXiv:2206.07682](https://arxiv.org/abs/2206.07682); 2022.
- [32] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. "Qlora: Efficient finetuning of quantized LMs," arXiv preprint. [arXiv:2305.14314](https://arxiv.org/abs/2305.14314); 2023.
- [33] Gehman S, Gururangan S, Sap M, Choi Y, Smith NA. "Realtotoxicityprompts: Evaluating neural toxic degeneration in language models," arXiv preprint. [arXiv:2009.11462](https://arxiv.org/abs/2009.11462); 2020.
- [34] Sheng E, Chang K-W, Natarajan P, Peng N. "The woman worked as a babysitter: On biases in language generation," arXiv preprint. [arXiv:1909.01326](https://arxiv.org/abs/1909.01326); 2019.
- [35] Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. "Large language models are zero-shot clinical information extractors," arXiv preprint. [arXiv:2205.12689](https://arxiv.org/abs/2205.12689); 2022.
- [36] Ouyang L, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Proces Syst* 2022;35:27730–44.
- [37] Jeblick K, et al. "ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports," arXiv preprint. [arXiv:2212.14882](https://arxiv.org/abs/2212.14882); 2022.
- [38] Wu C, Zhang X, Zhang Y, Wang Y, Xie W. "PMC-LLaMA: Further Finetuning LLaMA on Medical Papers," arXiv preprint. [arXiv:2304.14454](https://arxiv.org/abs/2304.14454); 2023.
- [39] Yang X, et al. A large language model for electronic health records. *npj Digital Medicine* 2022;5(1):194.
- [40] Luo R, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;23(6).
- [41] Singhal K, et al., "Large Language Models Encode Clinical Knowledge," arXiv preprint. [arXiv:2212.13138](https://arxiv.org/abs/2212.13138); 2022.
- [42] Chung HW, et al. "Scaling instruction-finetuned language models," arXiv preprint. [arXiv:2210.11416](https://arxiv.org/abs/2210.11416); 2022.
- [43] Singhal K, et al., "Towards Expert-Level Medical Question Answering with Large Language Models," arXiv preprint. [arXiv:2305.09617](https://arxiv.org/abs/2305.09617); 2023.
- [44] Kung TH, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2023;2(2):e0000198.
- [45] Jang D, Kim C-E. "Exploring the Potential of Large Language models in Traditional Korean Medicine: A Foundation Model Approach to Culturally-Adapted Healthcare," arXiv preprint. [arXiv:2303.17807](https://arxiv.org/abs/2303.17807); 2023.
- [46] A. Gilson et al., "How Well Does ChatGPT Do When Taking the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment," medRxiv, p. 2022.12. 23.22283901, 2022.
- [47] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," arXiv preprint [arXiv:2303.13375](https://arxiv.org/abs/2303.13375), 2023.
- [48] Wornow M, et al. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* 2023;6(1):135.
- [49] Moor M, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616(7956):259–65.
- [50] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2013, pp. 746–751.
- [51] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
- [52] M. E. Peters et al., "Deep Contextualized Word Representations," New Orleans, Louisiana, jun 2018: Association for Computational Linguistics, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237, doi:10.18653/v1/N18-1202. [Online]. Available: <http://aclanthology.org/N18-1202> [Online]. Available: doi:10.18653/v1/N18-1202.
- [53] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint [arXiv:1801.06146](https://arxiv.org/abs/1801.06146), 2018.
- [54] Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234–40.
- [55] E. Alsentzer et al., "Publicly available clinical BERT embeddings," arXiv preprint [arXiv:1904.03323](https://arxiv.org/abs/1904.03323), 2019.
- [56] Johnson AE, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3(1):1–9.
- [57] Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019;26(11):1297–304.
- [58] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017;5:135–46.
- [59] K. Huang, J. Altonaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," arXiv preprint [arXiv:1904.05342](https://arxiv.org/abs/1904.05342), 2019.
- [60] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform* 2014;47:1–10.
- [61] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;18(5):552–6.
- [62] Li J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016;2016.
- [63] Krallinger M, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Chem* 2015;7(1):1–17.
- [64] Smith L, et al. Overview of BioCreative II gene mention recognition. *Genome Biol* 2008;9(2):1–19.
- [65] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, 2004: Citeseer, pp. 70–75.
- [66] Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinform* 2010;11(1):1–17.
- [67] Pafilis E, et al. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One* 2013;8(6):e65390.
- [68] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinform* 2015;16(1):1–17.
- [69] Van Mulligen EM, et al. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *J Biomed Inform* 2012;45(5):879–84.
- [70] Krallinger M, et al. Overview of the BioCreative VI chemical-protein interaction Track. *Proceedings of the sixth BioCreative challenge evaluation workshop* 2017; 1:141–6.
- [71] Tsatsaronis G, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform* 2015;16(1):1–28.
- [72] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc* 2013;20(5):806–13.
- [73] Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. *J Biomed Inform* 2013;46:S5–12.
- [74] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," arXiv preprint [arXiv:1808.06752](https://arxiv.org/abs/1808.06752), 2018.
- [75] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;14(5):550–63.
- [76] Stubbs A, Kotfala C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *J Biomed Inform* 2015;58:S11–9.
- [77] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *J Biomed Inform* 2015;58:S20–9.
- [78] Suominen H, et al. Overview of the ShAre/CLEF eHealth evaluation lab 2013. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer; 2013. p. 212–31.
- [79] Kelly L, et al. Overview of the share/clef ehealth evaluation lab 2014. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer; 2014. p. 172–91.
- [80] S. Pradhan, N. Elhadad, W. W. Chapman, S. Manandhar, and G. Savova, "SemEval-2014 Task 7: Analysis of clinical text," in SemEval@ COLING, 2014, pp. 54–62.
- [81] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, and G. Savova, "SemEval-2015 task 14: Analysis of clinical text," in proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 303–310.
- [82] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen, "Semeval-2016 task 12: Clinical tempeval," in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1052–1062.
- [83] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets," arXiv preprint [arXiv:1906.05474](https://arxiv.org/abs/1906.05474), 2019.
- [84] Wang Y, et al. MedSTS: a resource for clinical semantic textual similarity. *Lang Resour Eval* 2020;54(1):57–72.
- [85] Sogancioglu G, Ozturk H, Ozgur A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 2017;33(14):i49–58.
- [86] Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform* 2013;46(5):914–20.
- [87] Baker S, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* 2016;32(3):432–40.
- [88] Gu Y, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 2021;3(1):1–23.
- [89] Zhou G, Su J. Exploring Deep Knowledge Resources in Biomedical Name Recognition, in BioNLP 2004, 2004/08// 2004, Geneva, Switzerland: COLING, pp. 99–102. [Online]. Available: <https://aclanthology.org/W04-1219>.
- [90] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," arXiv preprint [arXiv:1909.06146](https://arxiv.org/abs/1909.06146), 2019.
- [91] A. Nentidis, A. Krishnara, K. Bougiatiotis, G. Palioras, and I. Kakadiaris, "Results of the sixth edition of the BioASQ Challenge," Brussels, Belgium, November 2018: Association for Computational Linguistics, in Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering, pp. 1–10, doi:10.18653/v1/W18-5301. [Online]. Available: <https://aclanthology.org/W18-5301> [Online]. Available: doi:10.18653/v1/W18-5301.
- [92] Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc* 2020;27(12):1935–42.
- [93] Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J Am Med Inform Assoc* 2019;26(11):1163–71.
- [94] Henry S, Buchan K, Filannino M, Stubbs A, Uzuner Ö. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J Am Med Inform Assoc* 2020;27(1):3–12.
- [95] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692), 2019.
- [96] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint [arXiv:1909.11942](https://arxiv.org/abs/1909.11942), 2019.

- [97] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
- [98] Wei Q, et al. Relation extraction from clinical narratives using pre-trained language models. In: AMIA Annual Symposium Proceedings. 2019. American Medical Informatics Association; 2019. p. 1236.
- [99] Mayer T, Cabrio E, Villata S. "Transformer-based argument mining for healthcare applications," in ECAL. IOS Press 2020;2020:2108–15.
- [100] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.
- [101] K. Huang et al., "Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation," arXiv preprint arXiv:1912.11975, 2019.
- [102] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Proces Syst 2019;32.
- [103] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, "Biobert based named entity recognition in electronic medical record," in 2019 10th international conference on information technology in medicine and education (ITME), 2019: IEEE, pp. 49–52.
- [104] Alimova I, Tutubalina E. Multiple features for clinical relation extraction: A machine learning approach. J Biomed Inform 2020;103:103382.
- [105] Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019;42(1): 99–111.
- [106] Yoon W, Lee J, Kim D, Jeong M, Kang J. Pre-trained language model for biomedical question answering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2019. p. 727–40.
- [107] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," arXiv preprint arXiv:1606.05250, 2016.
- [108] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," arXiv preprint arXiv:1806.03822, 2018.
- [109] Ji Z, Wei Q, Xu H. BERT-based ranking for biomedical entity normalization. AMIA Summits on Translational Science Proceedings 2020;2020:269.
- [110] Pradhan S, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc 2015;22(1): 143–54.
- [111] Roberts K, Demner-Fushman D, Tonning JM. "Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track," in TAC. 2017.
- [112] Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y. Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models. JMIR Med Inform 2020;8(11):e19735.
- [113] Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. JMIR Med Inform 2020;8(11): e23375.
- [114] Zhang X, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019;132:103985.
- [115] S. Jiang, S. Zhao, K. Hou, Y. Liu, and L. Zhang, "A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition," in 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2019: IEEE, pp. 166–169.
- [116] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding," in 2020 International joint conference on neural networks (IJCNN), 2020: IEEE, pp. 1–8.
- [117] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records," in 2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei), 2019: IEEE, pp. 1–5.
- [118] Li X, Zhang H, Zhou X-H. Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inform 2020;107:103422.
- [119] Kim Y-M, Lee T-H. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020;20(7):1–9.
- [120] Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Appl Soft Comput 2020;97:106779.
- [121] R. Vunikili, H. Supriya, V. G. Marica, and O. Farri, "Clinical NER using Spanish BERT Embeddings," in IberLEF@ SEPLN, 2020, pp. 505–511.
- [122] Boudjellal N, et al. ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. Complexity 2021;2021:1–6.
- [123] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong, "Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus," arXiv preprint arXiv:2010.10391, 2020.
- [124] A. García-Pablos, N. Perez, and M. Cuadros, "Sensitive data detection and classification in Spanish clinical text: Experiments with BERT," arXiv preprint arXiv:2003.03106, 2020.
- [125] J. Mao and W. Liu, "Hadoken: a BERT-CRF Model for Medical Document Anonymization," in IberLEF@ SEPLN, 2019, pp. 720–726.
- [126] M. Marimon et al., "Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results," in IberLEF@ SEPLN, 2019, pp. 618–638.
- [127] Khan MR, Ziyadi M, AbdelHady M. "MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers," ed: arXiv. 2020.
- [128] R. Leaman and Z. Lu, "TaggerOne: joint named entity recognition and normalization with semi-Markov Models," (in eng), Bioinformatics (Oxford, England), vol. 32, no. 18, pp. 2839–2846, 2016/09/15/ 2016, doi:<https://doi.org/10.1093/bioinformatics/btw343>.
- [129] H.-L. Trieu, A.-K. D. Nguyen, N. Nguyen, M. Miwa, H. Takamura, and S. Ananiadou, "Coreference resolution in full text articles with bert and syntax-based mention filtering," in Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 196–205.
- [130] Cohen KB, et al. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. BMC Bioinform 2017;18(1):1–14.
- [131] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," arXiv preprint arXiv:1707.07045, 2017.
- [132] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes," (in en), J Biomed Inform, vol. 102, p. 103354, 2020/02/01/ 2020, doi:<https://doi.org/10.1016/j.jbi.2019.103354>.
- [133] O. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," (in eng), J Am Med Inform Assoc JAMIA, vol. 17, no. 5, pp. 519–523, 2010 2010, doi:<https://doi.org/10.1136/jamia.2010.004200>.
- [134] Lai P-T, Lu Z. BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer. Bioinformatics 2020;36(24):5678–85.
- [135] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W.-t. Yih, "Cross-sentence n-ary relation extraction with graph lstms," Trans Assoc Comput Linguist, vol. 5, pp. 101–115, 2017.
- [136] Wei C-H, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database 2016;2016.
- [137] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova, "A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 65–71.
- [138] Styler WF, et al. Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2014;2:143–54.
- [139] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee, "Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition," arXiv preprint arXiv:2010.03746, 2020.
- [140] Schmidt L, Weeds J, Higgins JPT. "Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks," ed: arXiv. 2020.
- [141] D. Jin and P. Szolovits, "PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks," in BioNLP 2018, 2018/07// 2018, Melbourne, Australia: Association for Computational Linguistics, pp. 67–75, doi:10.18653/v1/W18-2308. [Online]. Available: <https://aclanthology.org/W18-2308>.
- [142] Shen W, Wang J, Han J. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans Knowl Data Eng 2014;27(2):443–60.
- [143] Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019;7(3):e14830.
- [144] Xiong Y, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020;20(1):1–7.
- [145] Wang Y, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. Proceedings of the BioCreative/OHNLP Challenge, vol 2018;2018.
- [146] Z. Zhang, J. Liu, and N. Razavian, "BERT-XML: Large scale automated ICD coding using BERT pretraining," arXiv preprint arXiv:2006.03685, 2020.
- [147] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," Advances in Neural Information Processing Systems, vol. 32, 2019.
- [148] Biswas B, Pham T-H, Zhang P. TransICD: Transformer based code-wise attention model for explainable ICD coding. In: Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings. Springer; 2021. p. 469–78.
- [149] Z. Lin et al., "A structured self-attentive sentence embedding," arXiv preprint arXiv:1703.03130, 2017.
- [150] Cao K, Wei C, Gaidon A, Arechiga N, Ma T. Learning imbalanced datasets with label-distribution-aware margin loss. Adv Neural Inf Proces Syst 2019;32.
- [151] Wang Q, et al. A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes. J Biomed Inform 2020;105:103418.
- [152] G. López-García, J. M. Jerez, and F. J. Veredas, "ICB-UMA at CANTEMIST 2020: Automatic ICD-O Coding in Spanish with BERT," in IberLEF@ SEPLN, 2020, pp. 468–476.
- [153] G. López-García et al., "ICB-UMA at CLEF e-health 2020 task 1: Automatic ICD-10 coding in Spanish with BERT," in Proc. Work. Notes CLEF, Conf. Labs Eval. Forum, CEUR Workshop, 2020, pp. 1–15.
- [154] S. Remmer, A. Lamproudis, and H. Dalianis, "Multi-label diagnosis classification of Swedish discharge summaries–ICD-10 code assignment using KB-BERT," in International Conference Recent Advances in Natural Language Processing (RANLP'21), online, September 1–3, 2021, 2021: INCOMA Ltd., pp. 1158–1166.
- [155] K. Suvirat, D. Tanasanchonkul, K. Horsiritham, C. Kongkanol, T. Ingviya, and S. Chaichulee, "Automated Diagnosis Code Assignment of Thai Free-text Clinical Notes," in 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2022: IEEE, pp. 1–6.

- [156] S. Silvestri, F. Gargiulo, M. Ciampi, and G. De Pietro, "Exploit multilingual language model at scale for ICD-10 clinical text classification," in 2020 IEEE Symposium on Computers and Communications (ISCC), 2020: IEEE, pp. 1-7.
- [157] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.
- [158] Tubay B, Costa-Jussa MR. "Neural machine translation with the transformer and multi-source romance languages for the biomedical WMT 2018 task," in Proc Third Conference on Machine Translation: Shared Task Papers. 2018. p. 667-70.
- [159] A. Bérard, Z. M. Kim, V. Nikoulina, E. L. Park, and M. Gallé, "A multilingual neural machine translation model for biomedical data," arXiv preprint arXiv:2008.02878, 2020.
- [160] H. Liu, Y. Liang, L. Wang, X. Feng, and R. Guan, "BioNMT: A Biomedical neural machine translation system," INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL, vol. 15, no. 6, 2020.
- [161] X. Wang, Z. Tu, and S. Shi, "Tencent ai lab machine translation systems for the WMT21 biomedical translation task," in Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 874-878.
- [162] S. Subramanian, O. Hrinchuk, V. Adams, and O. Kuchaiev, "NVIDIA NeMo Neural Machine Translation Systems for English-German and English-Russian News and Biomedical Tasks at WMT21," arXiv preprint arXiv:2111.08634, 2021.
- [163] Choi E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. Proc AAAI Conference on artificial intelligence 2020;34(01):606-13.
- [164] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," Sci Data, vol. 5, p. 180178, Sep 11 2018, doi: <https://doi.org/10.1038/sdata.2018.178>.
- [165] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," arXiv preprint arXiv:1906.00346, 2019.
- [166] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang, "Sequential diagnosis prediction with transformer and ontological representation," in 2021 IEEE International Conference on Data Mining (ICDM), 2021: IEEE, pp. 489-498.
- [167] S. Darabi, M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "TAPER: Time-Aware Patient EHR Representation," IEEE J Biomed Health Inform, vol. 24, no. 11, pp. 3268-3275, Nov 2020, doi: <https://doi.org/10.1109/JBHI.2020.2984931>.
- [168] X. Luo et al., "Applying interpretable deep learning models to identify chronic cough patients using EHR data," Comput Methods Prog Biomed, vol. 210, p. 106395, Oct 2021, doi: <https://doi.org/10.1016/j.cmpb.2021.106395>.
- [169] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression," IEEE J Biomed Health Inform, vol. 25, no. 8, pp. 3121-3129, Aug 2021, doi: <https://doi.org/10.1109/JBHI.2021.3063721>.
- [170] Xu Z, So DR, Dai AM. Mufasa: Multimodal fusion architecture search for electronic health records. Proc AAAI Conf Artif Intel 2021;35(12):10532-40.
- [171] X. Zhang et al., "Learning robust patient representations from multi-modal electronic health records: a supervised deep learning approach," in Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), 2021: SIAM, pp. 585-593.
- [172] Y. Li et al., "BEHRT: Transformer for Electronic Health Records," Sci Rep, vol. 10, no. 1, p. 7155, Apr 28 2020, doi: <https://doi.org/10.1038/s41598-020-62922-y>.
- [173] E. Herrett et al., "Data Resource Profile: Clinical Practice Research Datalink (CPRD)," Int J Epidemiol, vol. 44, no. 3, pp. 827-36, Jun 2015, doi: <https://doi.org/10.1093/ije/dyv098>.
- [174] S. Rao et al., "An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure," IEEE J Biomed Health Inform, vol. 26, no. 7, pp. 3362-3372, Jul 2022, doi: <https://doi.org/10.1109/JBHI.2022.3148820>.
- [175] S. Rao et al., "Targeted-BEHTR: Deep Learning for Observational Causal Inference on Longitudinal Electronic Health Records," IEEE Trans Neural Netw Learn Syst, vol. PP, Jun 23 2022, doi: <https://doi.org/10.1109/TNNLS.2022.3183864>.
- [176] Y. Li et al., "Hi-BEHTR: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records," IEEE J Biomed Health Inform, vol. 27, no. 2, pp. 1106-1117, Feb 2023, doi: <https://doi.org/10.1109/JBHI.2022.3224727>.
- [177] L. Rasmussen, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," NPJ Digit Med, vol. 4, no. 1, p. 86, May 20 2021, doi: <https://doi.org/10.1038/s41746-021-00455-y>.
- [178] J. Luo, M. Ye, C. Xiao, and F. Ma, "Hitonet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 647-656.
- [179] H. Ren, J. Wang, W. X. Zhao, and N. Wu, "Rapt: Pre-training of time-aware transformer for learning robust healthcare representation," in Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 3503-3511.
- [180] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234-41.
- [181] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
- [182] B. Landman, Z. Xu, J. Iglesias, M. Styner, T. Langerak, and A. Klein, "Multi-atlas labeling beyond the cranial vault," URL: <https://www.synapse.org>, 2015.
- [183] Bernard O, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 2018;37(11):2514-25.
- [184] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," arXiv preprint arXiv:2102.10662, 2021.
- [185] Sirinukunwattana K, et al. Gland segmentation in colon histology images: The glas challenge contest. Med Image Anal 2017;35:489-502.
- [186] Kumar N, et al. A multi-organ nucleus segmentation challenge. IEEE Trans Med Imaging 2019;39(5):1380-91.
- [187] Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans Med Imaging 2017;36(7):1550-60.
- [188] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, "TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation," arXiv preprint arXiv:2107.05188, 2021.
- [189] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," arXiv preprint arXiv:2103.10504, 2021.
- [190] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv preprint arXiv:1902.09063, 2019.
- [191] Gao Y, Zhou M, Metaxas DN. UTNet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 61-71.
- [192] V. M. Campello et al., "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge," in IEEE Transactions on Medical Imaging, vol. 40, no. 12, pp. 3543-3554, Dec. 2021.
- [193] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," arXiv preprint arXiv:2102.08005, 2021.
- [194] Jha D, et al. Kvasis-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. Springer; 2020. p. 451-62.
- [195] Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99-111.
- [196] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imaging 2015;35(2):630-44.
- [197] Vázquez D, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Engi vol 2017;2017.
- [198] Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg 2014;9(2):283-93.
- [199] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," arXiv preprint arXiv:2103.03024, 2021.
- [200] H. Cao et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," arXiv preprint arXiv:2105.05537, 2021.
- [201] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation Transformer," arXiv preprint arXiv:2109.07162, 2021.
- [202] Z. Zhang, B. Sun, and W. Zhang, "Pyramid Medical Transformer for Medical Image Segmentation," arXiv preprint arXiv:2104.14702, 2021.
- [203] Andrzejczyk V, et al. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: 3D Head and Neck Tumor Segmentation in PET/CT Challenge. Springer; 2020. p. 1-21.
- [204] Ji Y, et al. Multi-Compound Transformer for Accurate Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 326-36.
- [205] Gamper J, Koobhanani NA, Benet K, Khuram A, Rajpoot N. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. Springer; 2019. p. 11-9.
- [206] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," arXiv preprint arXiv:1902.03368, 2019.
- [207] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "DS-TransUNet: Dual swin Transformer U-Net for medical image segmentation," arXiv preprint arXiv:2106.06716, 2021.
- [208] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. S. M. Goh, "Medical Image Segmentation using Squeeze-and-Expansion Transformers," arXiv preprint arXiv:2105.09511, 2021.
- [209] Orlando JI, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 2020;59:101570.
- [210] Sivaswamy J, Krishnadas S, Chakravarty A, Joshi G, Tabish AS. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomedical Imaging Data Papers 2015;2(1):1004.
- [211] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "RIM-ONE: An open retinal image database for optic nerve evaluation," in 2011 24th international symposium on computer-based medical systems (CBMS), 2011: IEEE, pp. 1-6.
- [212] B. Yun, Y. Wang, J. Chen, H. Wang, W. Shen, and Q. Li, "Spectr: Spectral transformer for hyperspectral pathology image segmentation," arXiv preprint arXiv:2103.03604, 2021.
- [213] Zhang Q, Li Q, Yu G, Sun L, Zhou M, Chu J. A multidimensional choledoch database and benchmarks for cholangiocarcinoma diagnosis. IEEE Access 2019;7:149414-21.
- [214] G. Xu, X. Wu, X. Zhang, and X. He, "LeViT-UNet: Make Faster Encoders with Transformer for Medical Image Segmentation," arXiv preprint arXiv:2107.08623, 2021.

- [215] Wang W, Chen C, Ding M, Yu H, Zha S, Li J. Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 109–19.
- [216] Menze BH, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 2014;34(10):1993–2024.
- [217] Bakas S, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* 2017;4(1):1–13.
- [218] B. Chen, Y. Liu, Z. Zhang, G. Lu, and D. Zhang, "TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation," arXiv preprint arXiv:2107.05274, 2021.
- [219] Shiraishi J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *Am J Roentgenol* 2000;174(1):71–4.
- [220] Jaeger S, Candemir S, Antani S, Wang Y-XJ, Lu P-X, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imaging Med Surg* 2014;4(6):475.
- [221] Y.-B. Tang, Y.-X. Tang, J. Xiao, and R. M. Summers, "Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation," in International Conference on Medical Imaging with Deep Learning, 2019: PMLR, pp. 457–467.
- [222] He X, et al. Benchmarking deep learning models and automated model design for COVID-19 detection with chest CT scans. *MedRxiv*. 2020.
- [223] Caicedo JC, et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nat Methods* 2019;16(12):1247–53.
- [224] Petit O, Thome N, Rambour C, Themyr L, Collins T, Soler L. U-net transformer: Self and cross attention for medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2021. p. 267–76.
- [225] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "AFTER-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation," arXiv preprint arXiv:2110.10403, 2021.
- [226] Chen X, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiother Oncol* 2021;160:175–84.
- [227] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, and D. Shen, "Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields," in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017: IEEE, pp. 1003–1006.
- [228] D. Guo and D. Terzopoulos, "A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation," in 2020 25th International Conference on Pattern Recognition (ICPR), 2021: IEEE, pp. 8857–8861.
- [229] Sun Q, Fang N, Liu Z, Zhao L, Wen Y, Lin H. HybridCTrm: Bridging CNN and Transformer for Multimodal Brain Image Segmentation. *J Healthc Eng* vol 2021;2021.
- [230] Mendrik AM, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comput Intell Neurosci* 2015;2015.
- [231] Wang L, et al. Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Trans Med Imaging* 2019;38(9):2219–30.
- [232] Y. Tang et al., "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022/06// 2022, New Orleans, LA, USA: IEEE, pp. 20698–20708, doi:10.1109/CVPR52688.2022.02007. [Online]. Available: <https://ieeexplore.ieee.org/document/9879123/>.
- [233] M. Antonelli et al., "The medical segmentation decathlon," arXiv preprint arXiv:2106.05735, 2021.
- [234] H. Zhang et al., "TiM-Net: Transformer in M-Net for Retinal Vessel Segmentation," (in en), *Journal of Healthcare Engineering*, vol. 2022, p. e9016401, 2022/07/11/ 2022, doi:<https://doi.org/10.1155/2022/9016401>.
- [235] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans Med Imaging*, vol. 19, no. 3, pp. 203–210, 2000/03// 2000, doi:<https://doi.org/10.1109/42.845178>.
- [236] C. G. Owen et al., "Measuring Retinal Vessel Tortuosity in 10-Year-Old Children: Validation of the Computer-Assisted Image Analysis of the Retina (CAIAR) Program," *Invest Ophthalmol Vis Sci*, vol. 50, no. 5, pp. 2004–2010, 2009/05/01/ 2009, doi:<https://doi.org/10.1167/iows.08-3018>.
- [237] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans Med Imaging*, vol. 23, no. 4, pp. 501–509, 2004/04// 2004, doi:<https://doi.org/10.1109/TMI.2004.825627>.
- [238] L. Wang, L. Yu, J. Zhu, H. Tang, F. Gou, and J. Wu, "Auxiliary Segmentation Method of Osteosarcoma in MRI Images Based on Denoising and Local Enhancement," (in en), *Healthcare*, vol. 10, no. 8, p. 1468, 2022/08// 2022, doi:<https://doi.org/10.3390/healthcare10081468>.
- [239] X. Shen, L. Wang, Y. Zhao, R. Liu, W. Qian, and H. Ma, "Dilated transformer: residual axial attention for breast ultrasound image segmentation," (in en), *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 9, pp. 4512–4528, 2022/09// 2022, doi:[10.21037/qims-22-33](https://doi.org/10.21037/qims-22-33).
- [240] Y. Zhang et al., "BUSIS: A Benchmark for Breast Ultrasound Image Segmentation," *Healthcare*, vol. 10, no. 4, p. 729, 2022/04/14/ 2022, doi:<https://doi.org/10.3390/healthcare10040729>.
- [241] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation," *IEEE Access*, vol. 10, pp. 80575–80586, 2022 2022, doi:<https://doi.org/10.1109/ACCESS.2022.3195241>.
- [242] Q. Jia and H. Shu, "BiTr-Unet: a CNN-Transformer Combined Network for MRI Brain Tumor Segmentation," arXiv preprint arXiv:2109.12271, 2021.
- [243] Kim H-I, Kim Y, Kim B, Shin DY, Lee SJ, Choi S-I. Hyoid bone tracking in a videofluoroscopic swallowing study using a deep-learning-based segmentation network. *Diagnostics* 2021;11(7):1147.
- [244] Zhang Z, Zhao T, Gay H, Zhang W, Sun B. Weaving attention U-net: A novel hybrid CNN and attention-based method for organs-at-risk segmentation in head and neck CT images. *Med Phys* 2021;48(11):7052–62.
- [245] B. Graham et al., "LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12259–12269.
- [246] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," arXiv preprint arXiv:2103.10504, 2021.
- [247] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
- [248] Xie Y, Zhang J, Shen C, Xia Y. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2021. p. 171–80.
- [249] G. M. Correia, V. Niculae, and A. F. Martins, "Adaptively sparse transformers," arXiv preprint arXiv:1909.00015, 2019.
- [250] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved Transformer for Volumetric Segmentation," arXiv preprint arXiv:2109.03201, 2021.
- [251] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv preprint arXiv:2103.14030, 2021.
- [252] Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. Springer; 2020. p. 108–26.
- [253] Karimi D, Vasylechko SD, Gholipour A. Convolution-free medical image segmentation using transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 78–88.
- [254] R. Mehta and J. Sivaswamy, "M-net: A Convolutional Neural Network for deep brain structure segmentation," in 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017/04// 2017, pp. 437–440, doi:10.1109/ISBI.2017.7950555. [Online]. Available: files/505/7950555.html.
- [255] A. Luthra, H. Sulakhe, T. Mittal, A. Iyer, and S. Yadav, "former: Edge Enhancement based Transformer for Medical Image Denoising," arXiv preprint arXiv:2109.08044, 2021.
- [256] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Proces Syst* 2021;34:12077–90.
- [257] Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. *Mach Vis Appl* 2020;31(1):1–18.
- [258] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration," arXiv preprint arXiv:2104.06468, 2021.
- [259] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV), 2016: IEEE, pp. 565–571.
- [260] Wang Y, Qian W, Zhang X. "A Transformer-based Network for Deformable Medical Image Registration," ed: arXiv. 2022.
- [261] Dosovitskiy A, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ed: arXiv. 2021.
- [262] Mok TCW, Chung ACS. "Affine Medical Image Registration with Coarse-to-Fine Vision Transformer," ed: arXiv. 2022.
- [263] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," (in eng) *J Cogn Neurosci*, vol. 19, no. 9, pp. 1498–1507, 2007/09// 2007, doi:<https://doi.org/10.1162/jocn.2007.19.9.1498>.
- [264] D. W. Shattuck et al., "Construction of a 3D probabilistic atlas of human cortical structures," (in eng) *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, 2008/02/01/ 2008, doi:<https://doi.org/10.1016/j.neuroimage.2007.09.031>.
- [265] G. v. Tulder, Y. Tong, and E. Marchiori, "Multi-view analysis of unregistered medical images using cross-view transformers," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021: Springer, pp. 104–113.
- [266] Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* 2017;4(1):1–9.
- [267] Irvin J, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc AAAI Conference on artificial intelligence* 2019;33(01):590–7.
- [268] J. Chen, Y. Du, Y. He, W. P. Segars, Y. Li, and E. C. Frey, "TransMorph: Transformer for unsupervised medical image registration," arXiv preprint arXiv:2111.10480, 2021.
- [269] "IXI Dataset," ed.
- [270] Y. Zhu and S. Lu, "Swin-VoxelMorph: A Symmetric Unsupervised Learning Model for Deformable Medical Image Registration Using Swin Transformer," L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., 2022 2022, Cham: Springer Nature Switzerland, in Lecture Notes in Computer Science, pp. 78–87, doi:10.1007/978-3-031-16446-0\_8.
- [271] C. R. Jack, Jr et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," *J Magn Reson Imaging*, vol. 27, no. 4, pp. 685–91, Apr 2008, doi:<https://doi.org/10.1002/jmri.21049>.

- [272] K. Marek et al., "The Parkinson Progression Marker Initiative (PPMI)," (in en), *Prog Neurobiol*, vol. 95, no. 4, pp. 629-635, 2011/12/01/ 2011, doi:<https://doi.org/10.1016/j.pneurobio.2011.09.005>.
- [273] Hou B, Kaassis G, Summers RM, Kainz B. RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 293–303.
- [274] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng, "Mimic-cxr database," PhysioNet <https://doi.org/10.13026/C2JT1Q>, 2019.
- [275] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," arXiv preprint [arXiv:1508.07909](https://arxiv.org/abs/1508.07909), 2015.
- [276] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.
- [277] Nicolson A, Dowling J, Koopman B. "AEHRC CSIRO in ImageCLEFmed Caption 2021," in CLEF2021 Working Notes, CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS. org; 2021.
- [278] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology Objects in COntext (ROCO): a multimodal image dataset," in Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: Springer, 2018, pp. 180-189.
- [279] Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med Image Anal* 2020;66: 101797.
- [280] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tinet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9049-9058.
- [281] P. Rajpurkar et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," arXiv preprint [arXiv:1712.06957](https://arxiv.org/abs/1712.06957), 2017.
- [282] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," (in en), *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021/01/01/ 2021, doi:10.1016/j.imu.2021.100557.
- [283] Rajpurkar P, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," ed: arXiv. 2017.
- [284] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," (in eng), *J Am Med Inform Assoc JAMIA*, vol. 23, no. 2, pp. 304-310, 2016/03// 2016, doi:<https://doi.org/10.1093/jamia/ocv080>.
- [285] You D, Liu F, Ge S, Xie X, Zhang J, Wu X. "AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation," ed: arXiv. 2022.
- [286] Johnson AEW, et al. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," ed: arXiv. 2019.
- [287] E. Pahwa, D. Mehta, S. Kapadia, D. Jain, and A. Luthra, "MedSkip: Medical Report Generation Using Skip Connections and Integrated Attention," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021 2021, pp. 3409-3415. [Online]. Available: [https://openaccess.thecvf.com/content/ICCV2021W/CVAMD/html/Pahwa\\_MedSkip\\_Medical\\_Report\\_Generation\\_Using\\_Skip\\_Connections\\_and\\_Integrated\\_Attention\\_ICCVW\\_2021\\_paper.html](https://openaccess.thecvf.com/content/ICCV2021W/CVAMD/html/Pahwa_MedSkip_Medical_Report_Generation_Using_Skip_Connections_and_Integrated_Attention_ICCVW_2021_paper.html).
- [288] B. Jing, P. Xie, and E. Xing, "On the Automatic Generation of Medical Imaging Reports," 2018 2018, pp. 2577-2586, doi:10.18653/v1/P18-1240. [Online]. Available: <http://arxiv.org/abs/1711.08195> [Online]. Available: files/340/1711.html.
- [289] Li M, Cai W, Verspoor K, Pan S, Liang X, Chang X. "Cross-modal Clinical Graph Transformer for Ophthalmic Report Generation," ed: arXiv. 2022.
- [290] M. Li et al., "FFA-IR: Towards an Explainable and Reliable Medical Report Generation Benchmark," in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021/10/31/ 2021. [Online]. Available: <https://openreview.net/forum?id=FgYTwJbjbf> [Online]. Available: files/343/forum.html.
- [291] F. Ren and Y. Zhou, "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," (in en), *IEEE Access*, vol. 8, pp. 50626-50636, 2020 2020, doi:<https://doi.org/10.1109/ACCESS.2020.2980024>.
- [292] "Visual Question Answering in the Medical Domain | ImageCLEF / LifeCLEF - Multimedia Retrieval in CLEF".
- [293] U. Naseem, M. Khushi, and J. Kim, "Vision-Language Transformer for Interpretable Pathology Visual Question Answering," *IEEE Journal of Biomedical and Health Informatics*, pp. 1-1, 2022 2022, doi:10.1109/JBHI.2022.3163751.
- [294] He X, Zhang Y, Mou L, Xing E, Xie P. "PathVQA: 30000+ Questions for Medical Visual Question Answering," ed: arXiv. 2020.
- [295] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multi-modal medical image synthesis," arXiv preprint [arXiv:2106.16031](https://arxiv.org/abs/2106.16031), 2021.
- [296] Goodfellow I, et al. Generative adversarial networks. *Commun ACM* 2020;63(11): 139-44.
- [297] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223-2232.
- [298] H. Wu et al., "Cvt: Introducing convolutions to vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22-31.
- [299] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, and S. A. Baker, "Vtgan: Semi-supervised retinal image synthesis and disease prediction using vision transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3235-3245.
- [300] Alipour S Hajeb Mohammad, Rabbani H, Akhlaghi MR. Diabetic retinopathy grading by digital curvelet transform. *Comput Math Methods Med* 2012;2012.
- [301] S. Yan, C. Wang, W. Chen, and J. Lyu, "Swin transformer-based GAN for multi-modal medical image translation," *Frontiers in Oncology*, vol. 12, 2022.
- [302] Hu Z, Liu H, Li Z, Yu Z. Data-Enabled Intelligence in Complex Industrial Systems Cross-Model Transformer Method for Medical Image Synthesis. *Complexity* 2021; 2021.
- [303] J. Liu, S. Pasumarthi, B. Duffy, E. Gong, G. Zaharchuk, and K. Datta, "One Model to Synthesize Them All: Multi-contrast Multi-scale Transformer for Missing Data Imputation," arXiv preprint [arXiv:2204.13738](https://arxiv.org/abs/2204.13738), 2022.
- [304] X. Zhang et al., "Ptnet: A high-resolution infant MRI synthesizer based on transformer," arXiv preprint [arXiv:2105.13993](https://arxiv.org/abs/2105.13993), 2021.
- [305] Makropoulou A, et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage* 2018;173:88-112.
- [306] K. Choromanski et al., "Rethinking attention with performers," arXiv preprint [arXiv:2009.14794](https://arxiv.org/abs/2009.14794), 2020.
- [307] Zhang X, et al. PTNet3D: A 3D High-Resolution Longitudinal Infant Brain MRI Synthesizer Based on Transformers. *IEEE Trans Med Imaging* 2022;41(10): 2925-40.
- [308] Howell BR, et al. The UNC/UMN Baby Connectome Project (BCP): An overview of the study design and protocol development. *NeuroImage* 2019;185:891-905.
- [309] Zhang Z, Yu L, Liang X, Zhao W, Xing L. TranST: dual-path transformer for low dose computed tomography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 55-64.
- [310] McCollough CH, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Med Phys* 2017;44(10):e339-52.
- [311] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17683-17693.
- [312] D. Wang, F. Fan, Z. Wu, R. Liu, F. Wang, and H. Yu, "CTformer: Convolution-free Token2Token Dilated Vision Transformer for Low-dose CT Denoising," arXiv preprint [arXiv:2202.13517](https://arxiv.org/abs/2202.13517), 2022.
- [313] Wang D, Wu Z, Yu H. Ted-net: Convolution-free t2t vision transformer-based encoder-decoder dilation network for low-dose ct denoising. In: International Workshop on Machine Learning in Medical Imaging. Springer; 2021. p. 416-25.
- [314] C. Wang, K. Shang, H. Zhang, Q. Li, Y. Hui, and S. K. Zhou, "Dudotrans: Dual-domain transformer provides more attention for sinogram restoration in sparse-view reconstruction," arXiv preprint [arXiv:2111.10790](https://arxiv.org/abs/2111.10790), 2021.
- [315] L. Yang and D. Zhang, "Low-Dose CT Denoising via Sinogram Inner-Structure Transformer," arXiv preprint [arXiv:2204.03163](https://arxiv.org/abs/2204.03163), 2022.
- [316] Jiayi Pan, Heye Zhang, Weifei Wu, Zhifan Gao, Weiwen Wu, Multi-domain integrative Swin transformer network for sparse-view tomographic reconstruction, *Patterns*, Volume 3, Issue 6, 2022.
- [317] Li R, et al. DDPTransformer: Dual-Domain With Parallel Transformer Network for Sparse View CT Image Reconstruction. *IEEE Transactions on Computational Imaging*. 2022.
- [318] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbeý and T. Çukur, "Unsupervised MRI Reconstruction via Zero-Shot Learned Adversarial Transformers," in *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1747-1763, July 2022.
- [319] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9446-9454.
- [320] Knoll F, et al. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiol Artif intel* 2020;2(1).
- [321] Feng C-M, Yan Y, Fu H, Chen L, Xu Y. Task transformer network for joint MRI reconstruction and super-resolution. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 307-17.
- [322] Feng C-M, et al. Multimodal Transformer for Accelerated MR Imaging," in: *IEEE Transactions on Medical Imaging* Oct. 2023;42(10):2804-16.
- [323] C. Fang, D. Zhang, L. Wang, Y. Zhang, L. Cheng, and J. Han, "Cross-Modality High-Frequency Transformer for MR Image Super-Resolution," arXiv preprint [arXiv:2203.15314](https://arxiv.org/abs/2203.15314), 2022.
- [324] P. Guo, Y. Mei, J. Zhou, S. Jiang, and V. M. Patel, "Reconformer: Accelerated mri reconstruction using recurrent transformer," arXiv preprint [arXiv:2201.09376](https://arxiv.org/abs/2201.09376), 2022.
- [325] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568-578.
- [326] G. Li et al., "Transformer-empowered Multi-scale Contextual Matching and Aggregation for Multi-contrast MRI Super-resolution," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20636-20645.
- [327] Gao C, Shih S-F, Finn JP, Zhong X. A Projection-Based K-space Transformer Network for Undersampled Radial MRI Reconstruction with Limited Training Subjects. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 726-36.
- [328] M. Ekanayake, K. Pawar, M. Harandi, G. Egan, and Z. Chen, "Multi-head Cascaded Swin Transformers with Attention to k-space Sampling Pattern for Accelerated MRI Reconstruction," arXiv preprint [arXiv:2207.08412](https://arxiv.org/abs/2207.08412), 2022.
- [329] Z. Zhao, T. Zhang, W. Xie, Y. Wang, and Y. Zhang, "K-Space Transformer for Fast MRI Reconstruction with Implicit Representation," arXiv preprint [arXiv:2206.06947](https://arxiv.org/abs/2206.06947), 2022.

- [330] K. Larsen, A. Pal, and Y. Rathi, "A Deep Learning Approach Using Masked Image Modeling for Reconstruction of Undersampled K-spaces," arXiv preprint arXiv:2208.11472, 2022.
- [331] Hu D, Zhang Y, Zhu J, Liu Q, Chen Y. TRANS-Net: Transformer-Enhanced Residual-Error AlterNative Suppression Network for MRI Reconstruction. *IEEE Trans Instrum Meas* 2022;71:1–13.
- [332] Z. Fabian and M. Soltanolkotabi, "HUMUS-Net: Hybrid unrolled multi-scale network architecture for accelerated MRI reconstruction," arXiv preprint arXiv:2203.08213, 2022.
- [333] J. Huang, Y. Wu, H. Wu, and G. Yang, "Fast MRI Reconstruction: How Powerful Transformers Are?," arXiv preprint arXiv:2201.09400, 2022.
- [334] C. Yan, G. Shi, and Z. Wu, "SMIR: A Transformer-Based Model for MRI super-resolution reconstruction," in 2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE), 2021: IEEE, pp. 1-6.
- [335] Huang J, Xing X, Gao Z, Yang G. Swin Deformable Attention U-Net Transformer (SDAUT) for Explainable Fast MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2022. p. 538–48.
- [336] B. Zhou et al., "Dsformer: A dual-domain self-supervised transformer for accelerated multi-contrast mri reconstruction," arXiv preprint arXiv:2201.10776, 2022.
- [337] Luo Y, et al. 3D transformer-GAN for high-quality PET reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 276–85.
- [338] Y. Fu et al., "A resource-efficient deep learning framework for low-dose brain PET image reconstruction and analysis," in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022: IEEE, pp. 1-5.
- [339] S.-I. Jang et al., "Spach Transformer: Spatial and Channel-wise Transformer Based on Local and Global Self-attentions for PET Image Denoising," arXiv preprint arXiv:2209.03300, 2022.
- [340] Yang Y, et al. A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer. *Am J Transl Res* 2021;13(2):743–56.
- [341] D. Ho, I. B. H. Tan, and M. Motani, "Predictive models for colorectal cancer recurrence using multi-modal healthcare data," in Proceedings of the Conference on Health, Inference, and Learning, 2021, pp. 204-213.
- [342] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," Advances in Neural Information Processing Systems, vol. 32, 2019.
- [343] E. Bacry, S. Gaiffas, A. Kabeshova, and Y. Yu, "ZiMM: a deep learning model for long term adverse events with non-clinical claims data," arXiv preprint arXiv:1911.05346, 2019.
- [344] Kabeshova A, Yu Y, Lukacs B, Bacry E, Gaiffas S. ZiMM: a deep learning model for long term and blurry relapses with non-clinical claims data. *J Biomed Inform* 2020;110:103531.
- [345] L. M. Scailteux et al., "French administrative health care database (SNDS): The value of its enrichment," *Therapie*, vol. 74, no. 2, pp. 215-223, Apr 2019, doi: <https://doi.org/10.1016/j.therap.2018.09.072>.
- [346] Zhang J, Nie Y, Chang J, Zhang JJ. Surgical Instruction Generation with Transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2021. p. 290–9.
- [347] E. Rojas-Muñoz, K. Couperus, and J. Wachs, "Daisi: Database for ai surgical instruction," arXiv preprint arXiv:2004.02809, 2020.
- [348] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311-318.
- [349] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74-81.
- [350] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65-72.
- [351] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566-4575.
- [352] Anderson P, Fernando B, Johnson M, Gould S. Spice: Semantic propositional image caption evaluation. In: European conference on computer vision. Springer; 2016. p. 382–98.
- [353] A. E. Aiello, A. Renson, and P. N. Zivich, "Social Media- and Internet-Based Disease Surveillance for Public Health," *Annu Rev Public Health*, vol. 41, pp. 101-118, Apr 2 2020, doi: <https://doi.org/10.1146/annurev-pubhealth-040119-094402>.
- [354] Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. *Am J Public Health* 2017;107(1):e1–8.
- [355] A. Mavragani, "Infodemiology and Infoveillance: Scoping Review," *J Med Internet Res*, vol. 22, no. 4, p. e16206, Apr 28 2020, doi: <https://doi.org/10.2196/16206>.
- [356] Abad ZSH, Butler GP, Thompson W, Lee J. Crowdsourcing for machine learning in public health surveillance: lessons learned from Amazon Mechanical Turk. *J Med Internet Res* 2022;24(1):e28749.
- [357] A. Breden and L. Moore, "Detecting adverse drug reactions from twitter through domain-specific preprocessing and bert ensembling," arXiv preprint arXiv:2005.06634, 2020.
- [358] S. Raval, H. Sedghamiz, E. Santus, T. Alhanai, M. Ghassemi, and E. Chersoni, "Exploring a Unified Sequence-To-Sequence Transformer for Medical Product Safety Monitoring in Social Media," arXiv preprint arXiv:2109.05815, 2021.
- [359] Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: Observational study. *JMIR Infodemiology* 2021;1(1):e26769.
- [360] Kummervold PE, et al. Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse. *JMIR Med Inform* 2021;9(10):e29584.
- [361] Alsudias L, Rayson P. Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study. *JMIR Med Inform* 2021;9(9):e27670.
- [362] Coleman JJ, Pontefract SK. Adverse drug reactions. *Clin Med* 2016;16(5):481.
- [363] Al Meslamani AZ. "Underreporting of Adverse Drug Events: A Look into the Extent, Causes and Potential Solutions," ed: Taylor & Francis. 2023.
- [364] D. Weissenbacher et al., "Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019," in Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, 2019, pp. 21-30.
- [365] Dirkson A, Verberne S, Sarker A, Kraaij W. Data-driven lexical normalization for medical social media. *Multimod Technol Interact* 2019;3(3):60.
- [366] A. Sakhovskiy, Z. Miftahutdinov, and E. Tutubalina, "KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects," in Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021, pp. 39-43.
- [367] Magge A, et al. "Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task," in Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task. 2021.
- [368] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," arXiv preprint arXiv:2010.09885, 2020.
- [369] Tutubalina E, Alimova I, Miftahutdinov Z, Sakhovskiy A, Malykh V, Nikolenko S. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* 2021;37(2):243–9.
- [370] Hussain S, Afzal H, Saeed R, Iltaf N, Umair MY. Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. *Comput Math Methods Med* 2021;2021.
- [371] Alvaro N, Miyao Y, Collier N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill* 2017;3(2):e6396.
- [372] Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc* 2017;24(4):813–21.
- [373] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform* 2012;45(5):885–92.
- [374] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019.
- [375] D. Weissenbacher, A. Sarker, M. Paul, and G. Gonzalez, "Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018," in Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, 2018, pp. 13-16.
- [376] X. Dai, S. Karimi, B. Hachey, and C. Paris, "An effective transition-based model for discontinuous NER," arXiv preprint arXiv:2004.13454, 2020.
- [377] Dietrich J, et al. Adverse events in twitter-development of a benchmark reference dataset: results from IMI WEB-RADR. *Drug Saf* 2020;43(5):467–78.
- [378] T. R. Goodwin, M. E. Savery, and D. Demner-Fushman, "Towards zero-shot conditional summarization with adaptive multi-task fine-tuning," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Conference on Empirical Methods in Natural Language Processing, 2020, vol. 2020: NIH Public Access, p. 3215.
- [379] M. Matero et al., "Suicide risk assessment with multi-level dual-context language and BERT," in Proceedings of the sixth workshop on computational linguistics and clinical psychology, 2019, pp. 39-44.
- [380] Kabir M, Ahmed T, Hasan M, Laskar M, Joarder TK, Mahmud H, Hasan K. DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities. *Comput Hum Behav* 2022;139:107503.
- [381] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
- [382] Ahne A, et al. Extraction of Explicit and Implicit Cause-Effect Relationships in Patient-Reported Diabetes-Related Tweets From 2017 to 2021: Deep Learning Approach. *JMIR Med Inform* 2022;10(7):e37201.
- [383] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," arXiv preprint arXiv:2005.10200, 2020.
- [384] Guidry JP, Jin Y, Orr CA, Messner M, Meganck S. Ebola on Instagram and Twitter: How health organizations address the health crisis in their social media engagement. *Public Relat Rev* 2017;43(3):477–86.
- [385] Reshi AA, et al. COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. *Healthcare* 2022;10, no. 3: MDPI:411.
- [386] Zhang M-L, Zhou Z-H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recogn* 2007;40(7):2038–48.
- [387] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2012, pp. 90-94.
- [388] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv preprint arXiv:2003.00104, 2020.

- [389] M. M. Islam and T. Iqbal, "Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020: IEEE, pp. 10285-10292.
- [390] Buffelli D, Vandin F. Attention-based deep learning framework for human activity recognition with user adaptation. *IEEE Sensors J* 2021;21(12):13474-83.
- [391] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in 2015 IEEE International conference on image processing (ICIP), 2015: IEEE, pp. 168-172.
- [392] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in 2012 IEEE computer society conference on computer vision and pattern recognition workshops, 2012: IEEE, pp. 20-27.
- [393] A. Kubota, T. Iqbal, J. A. Shah, and L. D. Riek, "Activity recognition in manufacturing: The roles of motion capture and sEMG+ inertial wearables in detecting fine vs. gross motion," in 2019 International Conference on Robotics and Automation (ICRA), 2019: IEEE, pp. 6533-6539.
- [394] A. Stisen et al., "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," in Proceedings of the 13th ACM conference on embedded networked sensor systems, 2015, pp. 127-140.
- [395] A. Reiss and D. Stricker, "Creating and benchmarking a new dataset for physical activity monitoring," in Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, 2012, pp. 1-8.
- [396] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in 2012 16th international symposium on wearable computers, 2012: IEEE, pp. 108-109.
- [397] M. Zhang and A. A. Sawchuk, "USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors," in Proceedings of the 2012 ACM conference on ubiquitous computing, 2012, pp. 1036-1043.
- [398] Sun Y, Shen Y, Ma L. MSST-RT: Multi-Stream Spatial-Temporal Relative Transformer for Skeleton-Based Action Recognition. *Sensors* 2021;21(16):5339.
- [399] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010-1019.
- [400] Liu J, Shahroudy A, Perez M, Wang G, Duan L-Y, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans Pattern Anal Mach Intell* 2019;42(10):2684-701.
- [401] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16266-16275.
- [402] S. Ramachandra, A. Hoelzemann, and K. Van Laerhoven, "Transformer Networks for Data Augmentation of Human Physical Activity Recognition," arXiv preprint arXiv:2109.01081, 2021.
- [403] Y. Tao et al., "Gated Transformer for Decoding Human Brain EEG Signals," Annu Int Conf IEEE Eng Med Biol Soc, vol. 2021, pp. 125-130, Nov 2021, doi:<https://doi.org/10.1109/EMBC46164.2021.9630210>.
- [404] Kostas D, Aroca-Ouellette S, Rudzicz F. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Front Hum Neurosci* 2021;15.
- [405] Obeid I, Picone J. The temple university hospital EEG data corpus. *Front Neurosci* 2016;10:196.
- [406] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000;101(23):e215-20.
- [407] Margaux P, Emmanuel M, Sébastien D, Olivier B, Jérémie M. Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling. *Advances in Human-Computer Interaction* 2012;2012(1):1-13. <https://doi.org/10.1155/2012/578295>.
- [408] Tangermann M, et al. Review of the BCI competition IV. *Front Neurosci* 2012;55.
- [409] Cai S, Li P, Su E, Xie L. Auditory Attention Detection via Cross-Modal Attention. *Front Neurosci* 2021;15:652058. <https://doi.org/10.3389/fnins.2021.652058>.
- [410] Fuglsang SA, Wong DD, Hjortkjær J. EEG and audio dataset for auditory attention decoding. Zenodo. 2018.
- [411] Fuglsang SA, Dau T, Hjortkjær J. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage* 2017;156:435-44.
- [412] P. Yi, K. Chen, Z. Ma, D. Zhao, X. Pu, and Y. Ren, "EEGDnet: Fusing Non-Local and Local Self-Similarity for 1-D EEG Signal Denoising with 2-D Transformer," arXiv preprint arXiv:2109.04235, 2021.
- [413] Zhang H, Zhao M, Wei C, Mantini D, Li Z, Liu Q. EEGdenoiseNet: a benchmark dataset for deep learning solutions of EEG denoising. *J Neural Eng* 2021;18(5):056057.
- [414] Behniaein B, Bhatti A, Rodenburg D, Hungler P, Etemad A. "A Transformer Architecture for Stress Detection from ECG," in: *Inte Symp Wearable Computers 2021;2021:132-4*.
- [415] H. Yu, T. Vaessen, I. Myin-Germeys, and A. Sano, "Modality Fusion Network and Personalized Attention in Momentary Stress Detection in the Wild," in 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), 2021: IEEE, pp. 1-8.
- [416] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in Proceedings of the 20th ACM international conference on multimodal interaction, 2018, pp. 400-408.
- [417] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 291-298.
- [418] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ECG signal processing and arrhythmia classification," *BMC Med Inform Decis Mak*, vol. 21, no. 1, p. 184, Jun 9 2021, doi:<https://doi.org/10.1186/s12911-021-01546-2>.
- [419] A. Khan and B. Lee, "Gene Transformer: Transformers for the Gene Expression-based Classification of Lung Cancer Subtypes," arXiv preprint arXiv:2108.11833, 2021.
- [420] N. Cancer Genome Atlas Research et al., "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat Genet*, vol. 45, no. 10, pp. 1113-20, Oct 2013, doi:<https://doi.org/10.1038/ng.2764>.
- [421] J. Clauwaert and W. Waegeman, "Novel transformer networks for improved sequence labeling in genomics," *IEEE/ACM Trans Comput Biol Bioinform*, vol. PP, Oct 30 2020, doi:<https://doi.org/10.1109/TCB.2020.3035021>.
- [422] A. Santos-Zavaleta et al., "RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12," *Nucleic Acids Res*, vol. 47, no. D1, pp. D212-D220, Jan 8 2019, doi:<https://doi.org/10.1093/nar/gky1077>.
- [423] F. Cunningham et al., "Ensembl 2019," *Nucleic Acids Res*, vol. 47, no. D1, pp. D745-D751, Jan 8 2019, doi:<https://doi.org/10.1093/nar/gky1113>.
- [424] P. Ye, Y. Luan, K. Chen, Y. Liu, C. Xiao, and Z. Xie, "MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing," *Nucleic Acids Res*, vol. 45, no. D1, pp. D85-D89, Jan 4 2017, doi:<https://doi.org/10.1093/nar/gkw950>.
- [425] Ettwiller L, Buswell J, Yigit E, Schildkraut I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* 2016;17(1):1-14.
- [426] B. Yan, M. Boitano, T. A. Clark, and L. Ettwiller, "SMRT-Cappable-seq reveals complex operon variants in bacteria," *Nat Commun*, vol. 9, no. 1, p. 3676, Sep 10 2018, doi:<https://doi.org/10.1038/s41467-018-05997-6>.
- [427] X. Ju, D. Li, and S. Liu, "Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria," *Nat Microbiol*, vol. 4, no. 11, pp. 1907-1918, Nov 2019, doi:<https://doi.org/10.1038/s41564-019-0500-z>.
- [428] J. Clauwaert, G. Menschaert, and W. Waegeman, "Explainability in transformer models for functional genomics," *Brief Bioinform*, vol. 22, no. 5, Sep 2 2021, doi: <https://doi.org/10.1093/bib/bbab060>.
- [429] Ji Y, Zhou Z, Liu H, Davuluri RV. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, Feb 4, 2021. <https://doi.org/10.1093/bioinformatics/btab083>.
- [430] J. Harrow et al., "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res*, vol. 22, no. 9, pp. 1760-74, Sep 2012, doi:<https://doi.org/10.1101/gr.135350.111>.
- [431] R. Drosos, G. Ambrosini, R. Cavin Perier, and P. Bucher, "EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D157-D64, Jan 2013, doi:<https://doi.org/10.1093/nar/gks1233>.
- [432] K. R. Rosenblloom et al., "ENCODE data in the UCSC Genome Browser: year 5 update," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D56-D63, Jan 2013, doi: <https://doi.org/10.1093/nar/gks1172>.
- [433] Serfling E, Jasen M, Schaffner W. Enhancers and eukaryotic gene transcription. *Trends Genet* 1985;1:224-30.
- [434] Liu B, Li K, Huang D-S, Chou K-C. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 2018;34(22):3835-42.
- [435] C. Jia and W. He, "EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features," *Sci Rep*, vol. 6, p. 38741, Dec 12 2016, doi:<https://doi.org/10.1038/srep38741>.
- [436] Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 2021;19:1750-8. <https://doi.org/10.1016/j.csbj.2021.03.022>.
- [437] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583-589, Aug 2021, doi:<https://doi.org/10.1038/s41586-021-03819-2>.
- [438] J. Jumper et al., "High accuracy protein structure prediction using deep learning," Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book), vol. 22, p. 24, 2020.
- [439] Q. Liu and L. Xie, "TransSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations," *PLoS Comput Biol*, vol. 17, no. 2, p. e1008653, Feb 2021, doi: <https://doi.org/10.1371/journal.pcbi.1008653>.
- [440] J. O'Neil et al., "An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies," *Mol Cancer Ther*, vol. 15, no. 6, pp. 1155-62, Jun 2016, doi:<https://doi.org/10.1158/1535-7163.MCT-15-0843>.
- [441] D. S. Wishart et al., "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res*, vol. 46, no. D1, pp. D1074-D1082, Jan 4 2018, doi: <https://doi.org/10.1093/nar/gkx1037>.
- [442] A. Gaulton et al., "The ChEMBL database in 2017," *Nucleic Acids Res*, vol. 45, no. D1, pp. D945-D954, Jan 4 2017, doi:<https://doi.org/10.1093/nar/gkw1074>.
- [443] Kim Y, Zheng S, Tang J, Jim Zheng W, Li Z, Jiang X. Anticancer drug synergy prediction in understudied tissues using transfer learning. *J Am Med Inform Assoc* 2021;28(1):42-51.
- [444] D. Zaikis and I. Vlahavas, "TP-DDI: Transformer-based pipeline for the extraction of Drug-Drug Interactions," *Artif Intell Med*, vol. 119, p. 102153, Sep 2021, doi: <https://doi.org/10.1016/j.artmed.2021.102153>.
- [445] Grechishnikova D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci Rep* 2021;11(1):1-13.
- [446] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and

- systems pharmacology," Nucleic Acids Res, vol. 44, no. D1, pp. D1045-53, Jan 4 2016, doi:<https://doi.org/10.1093/nar/gkv1072>.
- [447] Schwaller P, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci 2020;11(12):3316-25.
- [448] P. Schwaller et al., "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction," ACS Cent Sci, vol. 5, no. 9, pp. 1572-1583, Sep 25 2019, doi:<https://doi.org/10.1021/acscentsci.9b00576>.
- [449] Born J, et al. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. Machine Learning: Sci Technol 2021;2(2): 025024.
- [450] A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, A. Iuliano, and T. Laino, "Inferring experimental procedures from text-based representations of chemical reactions," Nat Commun, vol. 12, no. 1, p. 2573, May 6 2021, doi:<https://doi.org/10.1038/s41467-021-22951-1>.
- [451] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "MoTrans: Molecular Interaction Transformer for drug-target interaction prediction," Bioinformatics, vol. 37, no. 6, pp. 830-836, May 5 2021, doi:<https://doi.org/10.1093/bioinformatics/btaa880>.
- [452] Boutet E, Lieberherr D, Tognoli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot: Methods Mol Biol 2007;406:89-112. [https://doi.org/10.1007/978-1-59745-535-0\\_4](https://doi.org/10.1007/978-1-59745-535-0_4).
- [453] A. Gaulton et al., "ChEMBL: a large-scale bioactivity database for drug discovery," Nucleic Acids Res, vol. 40, no. Database issue, pp. D1100-7, Jan 2012, doi:<https://doi.org/10.1093/nar/gkr777>.
- [454] M. Zitnik, R. Sosic, and J. Leskovec, "BioSNAP Datasets: Stanford biomedical network dataset collection," Note: <http://snap.stanford.edu/biodata> Cited by, vol. 5, no. 1, 2018.
- [455] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," Nucleic Acids Res, vol. 35, no. Database issue, pp. D198-201, Jan 2007, doi:<https://doi.org/10.1093/nar/gkl999>.
- [456] M. I. Davis et al., "Comprehensive analysis of kinase inhibitor selectivity," Nat Biotechnol, vol. 29, no. 11, pp. 1046-51, Oct 30 2011, doi:<https://doi.org/10.1038/nbt.1990>.
- [457] M. Manica, A. Oskooei, J. Born, V. Subramanian, J. Saez-Rodriguez, and M. Rodriguez Martinez, "Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders," Mol Pharm, vol. 16, no. 12, pp. 4797-4806, Dec 2 2019, doi:<https://doi.org/10.1021/acs.molpharmaceut.9b00520>.
- [458] F. Iorio et al., "A Landscape of Pharmacogenomic Interactions in Cancer," Cell, vol. 166, no. 3, pp. 740-754, Jul 28 2016, doi:<https://doi.org/10.1016/j.cell.2016.06.017>.
- [459] P. Morris, R. St Clair, W. E. Hahn, and E. Barenholtz, "Predicting Binding from Screening Assays with Transformer Network Embeddings," J Chem Inf Model, vol. 60, no. 9, pp. 4191-4199, Sep 28 2020, doi:<https://doi.org/10.1021/acs.jcim.9b01212>.
- [460] S. Kim et al., "PubChem 2019 update: improved access to chemical data," Nucleic Acids Res, vol. 47, no. D1, pp. D1102-D1109, Jan 8 2019, doi:<https://doi.org/10.1093/nar/gky1033>.
- [461] Litsis EE, Das P, Kavraki LE. Prediction of drug metabolites using neural machine translation. Chem Sci 2020;11(47):12777-88.
- [462] Lowe DM. Extraction of chemical structures and reactions from the literature. University of Cambridge; 2012.
- [463] Wishart DS, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 2018;46(D1):D608-17.
- [464] Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res 2018;46(D1):D633-9.
- [465] E. Brunk et al., "Recon3D enables a three-dimensional view of gene variation in human metabolism," Nat Biotechnol, vol. 36, no. 3, pp. 272-281, Mar 2018, doi:<https://doi.org/10.1038/nbt.4072>.
- [466] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, and D. S. Wishart, "BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification," Aust J Chem, vol. 11, no. 1, p. 2, Jan 5 2019, doi:<https://doi.org/10.1186/s13321-018-0324-5>.
- [467] L. Ridder and M. Wagener, "SyGMA: combining expert knowledge and empirical scoring in the prediction of metabolites," ChemMedChem, vol. 3, no. 5, pp. 821-32, May 2008, doi:<https://doi.org/10.1002/cmdc.200700312>.
- [468] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 782-791.
- [469] M. Böhle, M. Fritz, and B. Schiele, "Holistically Explainable Vision Transformers," arXiv preprint [arXiv:2301.08669](https://arxiv.org/abs/2301.08669), 2023.
- [470] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021;4(1):1-13.
- [471] Li Y, et al. BEHRT: transformer for electronic health records. Sci Rep 2020;10(1): 1-12.
- [472] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," arXiv preprint [arXiv:1906.02243](https://arxiv.org/abs/1906.02243), 2019.
- [473] "AI and Compute," ed: OpenAI, 2018.
- [474] Schwartz R, Dodge J, Smith NA, Etzioni O. Green ai. Commun ACM 2020;63(12): 54-63.
- [475] Bloomfield P, Clutton-Brock P, Pencheon E, Magnusson J, Karpathakis K. Artificial Intelligence in the NHS: Climate and Emissions. J Clim Change Health 2021;4:100056.
- [476] Li C. Openai's gpt-3 language model: A technical overview. Blog Post. 2020.
- [477] Dodge J, et al. "Measuring the carbon intensity of ai in cloud instances," in 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1877-1894.
- [478] F. Lagunas, E. Charlaix, V. Sanh, and A. M. Rush, "Block pruning for faster transformers," arXiv preprint [arXiv:2109.04838](https://arxiv.org/abs/2109.04838), 2021.
- [479] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," arXiv preprint [arXiv:1908.09355](https://arxiv.org/abs/1908.09355), 2019.
- [480] Yao Z, Yazdani Aminabadi R, Zhang M, Wu X, Li C, He Y. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. Adv Neural Inf Proces Syst 2022;35:27168-83.
- [481] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?," arXiv preprint [arXiv:1905.10650](https://arxiv.org/abs/1905.10650), 2019.
- [482] K. Clark, U. Khanadelwali, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," arXiv preprint [arXiv:1906.04341](https://arxiv.org/abs/1906.04341), 2019.
- [483] Shen S, et al. Q-bert: Hessian based ultra low precision quantization of bert. Proc AAAI Conf Artif Intel 2020;34(05):8815-21.
- [484] Ganesh P, et al. Compressing Large-Scale Transformer-Based Models: A Case Study on BERT. Trans Assoc Comput Linguist 2021;9:1061-80. [https://doi.org/10.1162/tacl\\_a\\_00413](https://doi.org/10.1162/tacl_a_00413).
- [485] Zhao S, Gupta R, Song Y, Zhou D. Extreme language model compression with optimal subwords and shared projections. 2019.
- [486] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," arXiv preprint [arXiv:2111.00396](https://arxiv.org/abs/2111.00396), 2021.
- [487] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv preprint [arXiv:2312.00752](https://arxiv.org/abs/2312.00752), 2023.
- [488] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," arXiv preprint [arXiv:2307.08691](https://arxiv.org/abs/2307.08691), 2023.
- [489] R. Fletcher, A. Nakashima, and O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," vol. 3, ed: Frontiers Media SA, 2021, p. 561802.
- [490] S. Nerella, J. Cupka, M. Ruppert, P. Tighe, A. Bihorac, and P. Rashidi, "Pain Action Unit Detection in Critically Ill Patients," in 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021: IEEE, pp. 645-651.
- [491] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464): 447-53.
- [492] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: quantifying biases in clinical contextual word embeddings," in proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 110-120.
- [493] Nazer LH, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023;2(6):e0000278.
- [494] Ji Z, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1-38.
- [495] R. Ngo, "The alignment problem from a deep learning perspective," arXiv preprint [arXiv:2209.00626](https://arxiv.org/abs/2209.00626), 2022.
- [496] Gostin LO. National health information privacy: regulations under the Health Insurance Portability and Accountability Act. JAMA 2001;285(23):3015-21.
- [497] Van Panhuis WG, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014;14(1):1-9.
- [498] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021;5(1):1-19.
- [499] Y. Kim, J. Sun, H. Yu, and X. Jiang, "Federated tensor factorization for computational phenotyping," in Proceedings of the 23rd ACM SIGKDD International conference on knowledge discovery and data mining, 2017, pp. 887-895.
- [500] Lee J, Sun J, Wang F, Wang S, Jun C-H, Jiang X. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Med Inform 2018;6(2):e7744.
- [501] Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform 2019;99:103291.
- [502] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "Braintorrent: A peer-to-peer environment for decentralized federated learning," arXiv preprint [arXiv:1905.06731](https://arxiv.org/abs/1905.06731), 2019.
- [503] Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer; 2019. p. 92-104.
- [504] Li W, et al. Privacy-preserving federated brain tumour segmentation. In: International workshop on machine learning in medical imaging. Springer; 2019. p. 133-41.
- [505] Rieke N, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3(1):1-7.