SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Transformer-Based Natural Language Processing Model for Automated Clinical Diagnosis and Electronic Health Record Summarization

Rashmi Jain[1], Dr. Shikhar Verma[2], Dr. Jimmy Narayan[3], Manish M Goswami[4], Sachin U. Balvir[5], Anup Gade[6], Dr. Komal Patel[7], Pushpalatha P[8]

[1]Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra-441501, India. Email: rashmilalitjain@gmail.com
[2]Professor , MSOPS, Maharishi University of Information Technology, Lucknow, Uttar Pradesh, India, Email: shikhar.verma@muit.in, Orcid Id- https://orcid.org/0000-0002-2481-395X
[3]Professor, Department of Gastroenterology, IMS and SUM Hospital, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, Email: jimmynarayan@soa.ac.in, Orcid Id- 0000-0002-8451-1558
[4]Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra-441501, India. Email: manish.goswami@sbjit.edu.in
[5]Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra-441501, India. Email: sachin_balvir@yahoo.com
[6]Department of Computer Science and Engineering, S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra-441501, India. Email: gadeanup@gmail.com
[7]Consultant, Department of Gynaecology, Parul University, PO Limda, Tal. Waghodia, District Vadodara, Gujarat, India, Email: drkomal@paruluniversity.ac.in
[8]Computer Science, Assistant Professor, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India, Email: pushpalathap@maher.ac.in

Abstract

Electronic Health Records (EHRs) hold vast amounts of unstructured clinical notes that would be hard to efficiently analyze within traditional healthcare information systems. Correct interpretation of the clinical notes is crucial for prompt diagnosis, adequate treatment planning, and decrease in physician documentation workload. Yet clinical texts may present certain terminology that is ambiguous, some abbreviations, and some different writing styles, all of which pose challenges to automated medical text processing. This research aims to develop an automated clinical diagnosis prediction framework and an EHR summarization framework based on Natural Language Processing (NLP) and transformer. The proposed model uses contextual embedding, multi-head attention mechanisms and multitask learning to model disease classification and to summarize a patient's raw records into a concise description. Benchmark healthcare datasets such as the Medical Information Mission to Intelligent City (MIMIC-III) clinical notes were used for experimentation, and the clinical notes were preprocessed using tokenization, normalization, and medical entity extraction. The framework was assessed on a set of diagnosis prediction metrics (accuracy, precision, recall, F1-score, and AUROC) as well as summarization metrics (ROUGE and BLEU scores). Experimental results show that the model outperforms the current deep learning models in terms of the increased accuracy of prediction and the quality of summarization. The proposed framework has good potential for intelligent clinical decision support and healthcare documentation automation.

Keywords: Transformer-based NLP, Clinical diagnosis prediction, Electronic health records, Medical text summarization, Healthcare informatics, Deep learning in healthcare

1. Introduction

The digitization of healthcare systems has advanced quickly, increasing the use of an electronic health record (EHR) to centralize patient information and integrate clinical systems and workflows (Johnson et al., 2016; Johnson et al., 2023). With the growing amount of clinical data being created in the hospital or health care environment, there is a great opportunity for the use of intelligent data-driven healthcare analytics (Miotto et al. 2018; Rajkomar et al. 2019). Most of the data within EHR is, however, unstructured and heterogeneous,

making for manual interpretation that is time consuming and computationally difficult. Documentation requirements for clinicians can be heavy while verifying and accessing patient information for efficient clinical decision making is crucial. Thus, automated systems for healthcare text analysis have emerged as a key research focus in medical artificial intelligence (AI) (Shickel et al., 2017).

Artificial Intelligence (AI) and Natural Language Processing (NLP) have recently been successful at enhancing the ability of machines to process complex medical terminology and clinical context (Miotto et al., 2018; Rajkomar et al., 2019). Specifically, transformer-based deep learning architectures have been found to perform exceptionally well in a variety of language understanding tasks for their ability to 'embed' words in a context as well as learn 'attention' to other words (Vaswani et al., 2017; Devlin et al., 2019). BERT, BioBERT and ClinicalBERT models are among the several that have achieved good performance in healthcare text summarization, clinical prediction, medical entity recognition, and disease classification (Devlin et al., 2019; Lee et al., 2020; Huang et al., 2019; Alsentzer et al., 2019). However, the current methods have been developed more towards the single task of automated diagnosis prediction or EHR summarization without leveraging other components, which is not effective in the integrated clinical decision support environment.

A major drawback of current clinical NLP systems is their lack of the ability to handle long-range semantic dependency or context relationships in long medical narratives (Vaswani et al., 2017; Peng et al., 2019). Traditional machine learning algorithms like linear regression, support vector machines, and neural networks are less scalable, have limited contextual information and poor generalization across different datasets relevant to healthcare (Shickel et al., 2017; Miotto et al., 2018). Furthermore, there is a need for multitask transformer frameworks that are clinically reliable and can perform diagnosis prediction and summarize medical records in a concise form at the same time (Lewis et al., 2020; Raffel et al., 2020).

To cope with these challenges, this study suggests an NLP-based automated clinical diagnosis prediction and EHR summarization approach based on transformer. In the proposed model, contextual attention mechanisms and multitask learning are combined to enhance the understanding of clinical text and healthcare information extraction (Vaswani et al., 2017; Devlin et al., 2019). The framework is tested via standard benchmark clinical datasets and multiple performance metrics to validate the effectiveness of the framework against the baseline methods that exist (Johnson et al., 2016; Johnson et al., 2023).

2. Related Work

The ability of NLP to process vast amounts of unstructured clinical text to gain insight from medical information has made it a large research field in the healthcare sector (Miotto, et al., 2018; Rajkomar, et al., 2019). Clinical text mining techniques have been widely used in the field of healthcare analytics and clinical decision support, particularly in the context of electronic health records (EHRs) and physician notes, discharge summaries, pathology reports, and radiology interpretations, among others (Shickel et al., 2017; Kormilitzin et al., 2021). The first generation of NLP based medical systems focused mainly on rule based methods and were based on handcrafted linguistic features in the context of medical concept detection and patient information extraction. Clinical narratives were commonly processed using medical entity recognition methods to extract various entities such as diseases, medications, symptoms, procedures and laboratory findings (Kormilitzin et al., 2021). Thus, these techniques played a significant role in supporting automated healthcare documentation and information retrieval systems but care needs to be taken to ensure that they can handle the vocabulary variations in specific medical contexts and the general ambiguities found in medical text.

In healthcare NLP tasks, transformer-based deep learning models greatly enhanced the understanding of medical text and the representation of context (Devlin et al., 2019; Vaswani et al., 2017). Compared to conventional recurrent neural networks (RNNs) that were the predominant choice for studying long-range dependencies and semantic relationships in textual information, transformer architectures are better able to capture these key features to a greater degree (Vaswani et al., 2017). BERT, or Bidirectional Encoder Representations from Transformers, exhibited great performance in general NLP tasks and was adapted to several healthcare-specific uses, such as developing an approach to the specific task of Covid-19 detection (Devlin et al., 2019). By leveraging extensive biomedical text collections, BioBERT further extended the BERT

model and achieved better results in biomedical text mining and medical entity recognition. (Lee et al., 2020). Over the past few years, clinical language understanding has been improved through an adaptation of BERT on electronic health records and hospital notes, yielding improved clinical prediction and patient outcome analysis (Huang et al., 2019; Alsentzer et al., 2019). In addition, a large number of GPT-based transformer models have also been used in the field of medical report generation, healthcare question answering and clinical text summarization recently because of its advanced generative language modelling capabilities (Lewis et al., 2020; Raffel et al., 2020; Jin et al., 2021).

Clinical diagnosis systems have now gone beyond the rule-based expert systems to high-level machine learning and deep learning systems (Rajkomar et al., 2019). In early diagnosis systems, manually-designed medical rules and trees were utilized to assist the identification and treatment recommendations of diseases. Later, machine learning techniques such as Support Vector Machines, Random Forests, and Naive Bayes classifiers were developed that learned from clinical datasets to better predict the accuracy. Recently, the application of deep learning models which employ Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and transformer architecture have demonstrated high performance in the prediction of various diseases and the detection of complex clinical patterns and contextual dependencies (Shickel et al., 2017; Devlin et al., 2019; Huang et al., 2019).

Moreover, summarization of EHR is also a crucial NLP application in healthcare that seeks to alleviate doctors' burdens and make patient information more accessible (Lewis et al., 2020; Raffel et al., 2020). Summarizing methods can be roughly divided into two categories: extractive and abstractive. Extractive summarization methods are those that pick out key sentences from clinical documents; abstractive summarization methods are those that produce a shorter summary based on a semantic understanding and the use of language generation mechanisms. Recent transformer-based summarization models proved to be more successful with respect to contextual coherence and semantic preservation than the conventional sequence-to-sequence models (Vaswani et al., 2017; Lewis et al., 2020).

However, there are certain constraints in prevalent NLP-based studies in the field of health care. Most existing methods concentrate on the separate tasks of predicting diagnosis or summarizing the EHRs without proposing unified multitask learning frameworks. Moreover, traditional approaches have suffered from difficulties in dealing with the contextually complex meaning of lengthy clinical narratives and have not been well generalized to other healthcare datasets (Peng et al., 2019; Kormilitzin et al., 2021). However, challenges remain in the deployment of transformer-based healthcare NLP systems to real-world medical settings, such as being scalable, interpretable, and clinically reliable (Miotto et al., 2018; Rajkomar et al., 2019).

3. Materials and Methods

3.1 Clinical Dataset and Data Preprocessing

The framework of clinical NLP proposed here was built and tested with the benchmark publicly available healthcare datasets which are commonly used in the medical language processing research. The rich source of patient perspectives on their own electronic health records and annotated clinical narratives was a consideration of clinical data sets including the MIMIC-III, MIMIC-IV, and i2b2 clinical corpus. These records feature a variety of data about patients' health, such as physician notes, discharge summaries, information on medications, diagnosis records, lab observations, and patient treatment descriptions. These structured diagnosis labels and unstructured clinical annotations are suitable for both the automated diagnosis prediction and EHR summarization tasks.

The MIMIC-III and MIMIC-IV are large-scale, de-identified intensive care unit (ICU) data collected from hospital information systems (HIS), which include longitudinal patient information and associated ICD diagnosis codes. These datasets are full of admissions notes from patients and extensively used for NLP benchmarking in the healthcare sector. The i2b2 clinical dataset contains three types of annotated medical narratives: those specifically annotated for clinical information extraction and those annotated for clinical NLP evaluation. By leveraging these datasets, the proposed framework will gain the ability to acquire a range of medical terms,

context-specific clinical patterns, and disease-related semantic relationships, thereby supporting it in handling various healthcare scenarios.

To guarantee the quality of the preprocessing operations and to aid in efficient language modeling using the transformer, several preprocessing steps were taken before the models were used to train. To ensure the privacy of patient information and adhere to healthcare data protection regulations, the initial step involved de-identifying all patient-related personally identifiable information (PII). Typographical errors, symbols repetitions, missing sentences and irrelevant characters are quite common in clinical text; hence noise removal methods were used in order to remove the useless textual noise from the dataset.

Then, the medical abbreviation normalization was carried out to translate the medical abbreviations in the clinical documents into the standardized medical expressions, in order to make clinical documents more semantically coherent. Clinical narratives were subsequently segmented into meaningful text segments for transformer embedding generation using tokenization techniques.Using the tokenization techniques, clinical narratives were then tokenized into meaningful text segments, which can be used for embedding generation by the transformer. To avoid the computational redundancy and make feature learning more efficient, common stop words with small semantic contribution were filtered. Furthermore, clinical entity extraction techniques were used to identify clinical concepts and relationships including diseases, medications, symptoms, laboratory findings and procedures from clinical narratives.

ICD code mapping was used to correlate free-text with the standard labels of disease classification, for automated diagnosis prediction. Finally, the obtained results were further split into training, validation, and test sets to assess the model unbiasedly and to evaluate the model-generalization. This pre-processing pipeline improved the quality of the clinical representations inputted to the NLP model and the proposed transformer based healthcare NLP model worked better as a result.

3.2 Proposed Transformer-Based Clinical NLP Framework

A multitask deep learning architecture is proposed to accomplish two tasks of clinical NLP: the clinical diagnosis prediction and electronic health record (EHR) summarization, with a transformer-based framework. The architecture incorporates contextual language understanding, transformer-based attention mechanisms, and dual-task learning to develop better semantic interpretation of clinical narratives and to better support making health care decisions. Figure 1 above shows the overall architecture of the proposed framework.

The system takes in raw EHR data such as clinical notes, discharge summaries, laboratory reports, patient history records, etc.The system first receives the raw EHR data, which include patient history, clinical notes, laboratory reports, and discharge summaries. Clinical narratives are not necessarily consistent in formatting, the use of abbreviations, and in irrelevant textual information, for this reason a preprocessing module is used to improve the data quality and the semantic consistency. The preprocessing stage carries out de-identification, noise removal, medical abbreviation normalization, tokenization, ICD code mapping and splits the dataset for transformer-based representation learning of the clinical text.

The preprocessed clinical text is then sent to the embedding layer, which produces embeddings for the tokens and positional embeddings to retain the meaning and order in the medical texts. These embedded representations are then passed through a series of transformer encoded blocks with multiple attention heads, feed-forward neural networks, residual connections and normalization layers. This transformer encoder is used to capture long-range contexts and complex semantics in the clinical documents.

The attention mechanism also facilitates contextual feature extraction by calculating inter-dependency relations between query, key and value representations. The framework can extract clinically relevant patterns and features in the text more accurately by using attention-guided learning. The learned contextual representations are then mapped to a common feature representation layer for both diagnosis prediction and a summarization of the EHRs.

In the case of automated diagnosis prediction, the diagnosis head carries out a multi-label disease classification using sigmoid and softmax activation layers, and provides ICD disease labels. In the training phase, it uses

cross-entropy optimization for improving classification accuracy and also the predictive reliability. At the same time, the summarization decoder adopts context-aware transformer decoding and sequence generation mechanisms to generate abstractive summaries of long clinical narratives, which are brief and retain key medical information and semantic coherence. Lastly, the framework can create two outputs: algorithmic clinical diagnosis predictions and brief EHR summaries. The multitask transformer-based framework enhances the understanding and analysis of healthcare documents, facilitates clinical decision support, and boosts the efficiency of medical documentation.

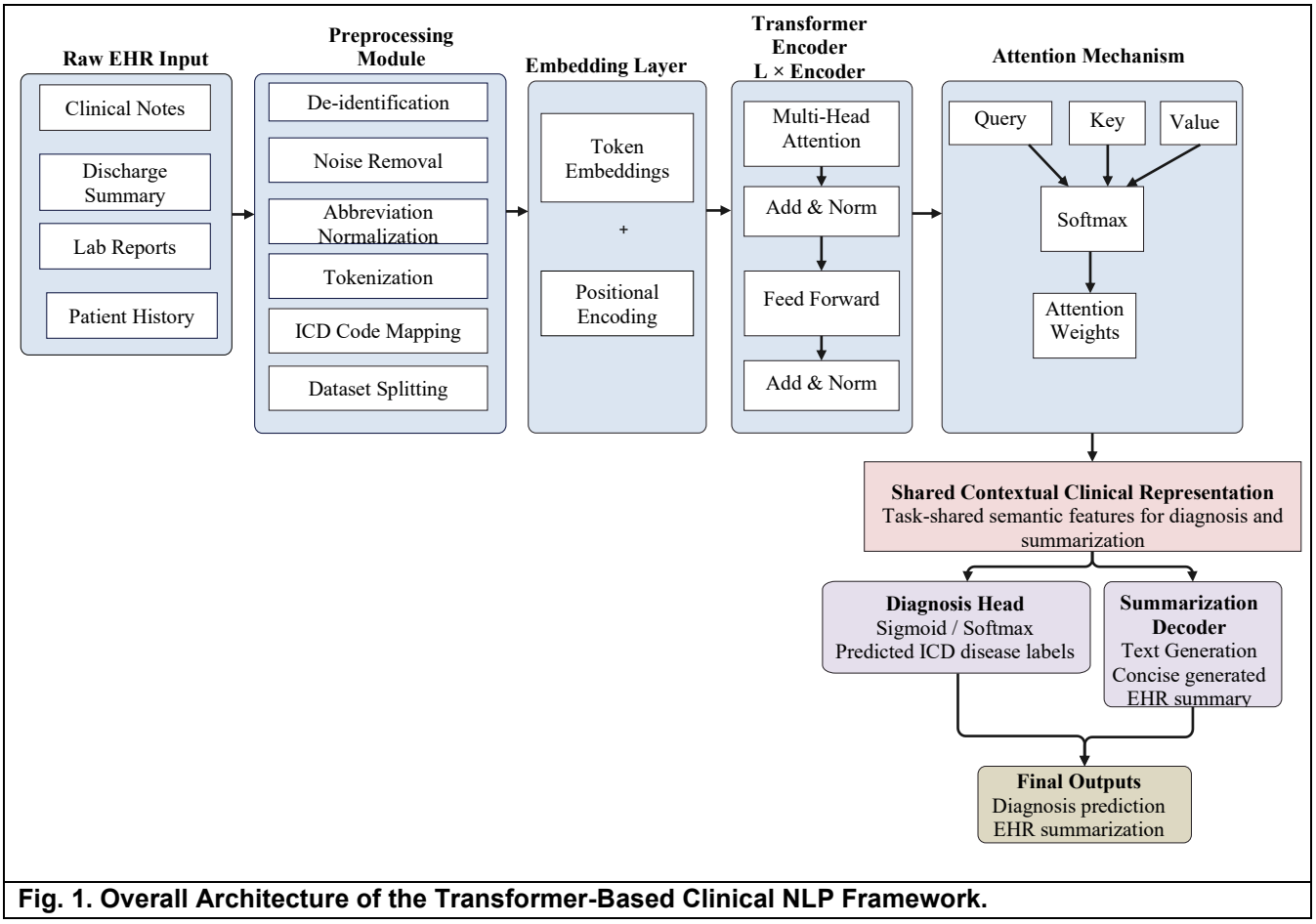

Fig. 1. Overall Architecture of the Transformer-Based Clinical NLP Framework.

The proposed framework comprises the raw input data of EHRs, preprocessing with operations, contextual representation learning module using transformer encoder network, attention guided feature extraction module, diagnosis prediction head, summarization decoder, and ultimate multi task healthcare output.

3.3 Training Strategy and Evaluation Metrics

A supervised multi-task learning strategy was adopted for training the proposed clinical NLP framework, which serves not only for automated diagnosis prediction task but also for EHR summarization. To assess the performance in an un-biased manner, the dataset was split into three sets: 70% training, 15% validation, and 15% testing. The transformer model was optimized using the AdamW optimizer as it was found to be effective in handling weight decay and generalized the transformer model. To avoid the unstable convergence, a learning rate scheduling strategy was used for the gradual change of learning rate during the training process.To adjust the learning rate smoothly in the training process, the unstable convergence phenomenon is prevented by using the learning rate scheduling strategy.

The model was trained with the batch size of 16 or 32 (depending on the available GPU memory), and the maximum sequence length of 512 tokens for clinical notes. Training was done for 10-20 epochs and early stopping was employed at the point of non-reduction of validation loss for trailing epochs. To prevent

overfitting, dropout regularization with a value of 0.1–0.3 was applied. The cross-entropy loss was used in the diagnosis prediction and the sequence-generation loss was used in the summarization decoder. The overall clinical text processing, multitask learning and performance evaluation process is illustrated in Figure 2. The accuracy, precision, recall, F1-score and AUROC metrics were used to assess the performance of the model for diagnosis prediction. These metrics were chosen because the prediction of clinical diagnosis based on the data needs to be accurate and be highly sensitive to the disease conditions. The summaries generated were assessed for EHR summarization using ROUGE-1, ROUGE-2, ROUGE-L, BLEU score and BERTScore. The word overlap and phrase overlap metrics (ROUGE), the linguistic similarity metric (BLEU), and the semantic preservation metric (BLERTS with BERT embeddings) are all measures of the ROUGE metrics.

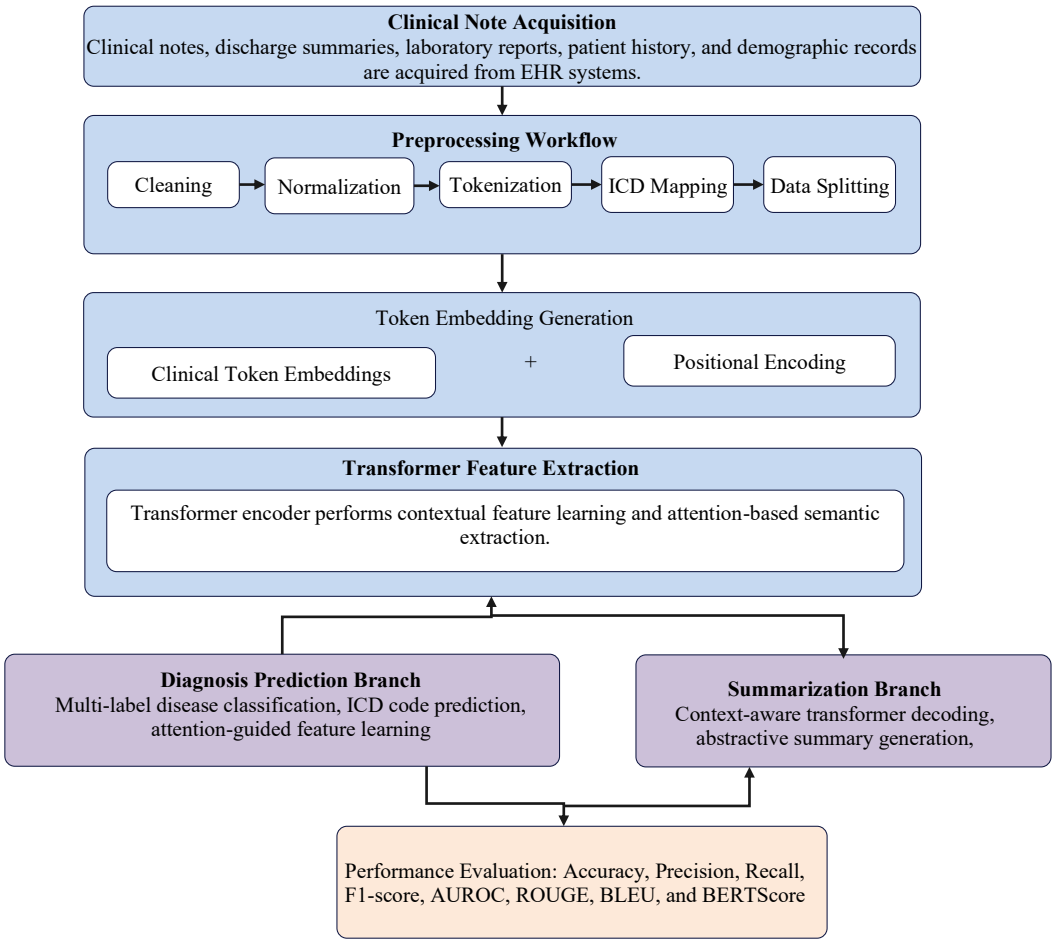

Fig. 2. Workflow of Clinical Text Processing and Multi-Task Learning Pipeline.

The figure provides a comprehensive overview of the entire process, starting from the acquisition of clinical notes, preprocessing, generation of token embeddings, feature extraction using the transformer architecture, prediction of diagnosis, summarization of EHR, and evaluation of the performance.

4. Results and Analysis

4.1 Predict the performance of the diagnosis.

The results of the proposed transformer-based clinical NLP framework were then compared with various baseline deep learning models such as LSTM, BiLSTM, CNN-LSTM, ClinicalBERT, and BERT for the diagnosis prediction. Standard clinical classification metrics (accuracy, precision, recall, F1-score, AUROC) were used to evaluate the predictive reliability and contextual understanding capability of each model comprehensively. Comparative performance results are shown in Table 1.

Table 1. Performance Comparison of Diagnosis Prediction Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUROC (%)
LSTM	84.21	82.45	81.76	82.10	86.10
BiLSTM	86.37	84.92	84.15	84.53	88.40
CNN-LSTM	88.14	86.73	85.96	86.34	90.30
BERT	90.52	89.11	88.64	88.87	92.60
ClinicalBERT	92.08	91.27	90.84	91.05	94.70
Proposed Transformer Model	94.36	93.81	93.24	93.52	97.20

Table 1 indicates that conventional recurrent neural network models like LSTM and BiLSTM showed relatively poor diagnosis prediction performance because of their lower capacity to represent long-range contextual information in intricate clinical narratives. The CNN-LSTM model was developed by integrating local feature extraction with sequential learning mechanisms to enhance its performance in this study. Contextual embedding and self-attention learning capabilities of transformer-based architectures, such as BERT and ClinicalBERT, had led to considerable improvements in comparison with using traditional deep learning approaches.

The proposed NLP model based on Transformer shed the highest overall performance for all the evaluation metrics, including accuracy (94.36%), precision (93.81%), recall (93.24%), F1-score (93.52%) and AUROC (97.20%). The contextual transformer encoding, attention-guided feature extraction, and shared semantic representation by multitask learning contribute to the improvement in performance. The model demonstrated excellent performance on the challenge task of identifying clinically relevant disease patterns and the ability to make reliable diagnosis predictions on unstructured EHR. The curves of diagnosis prediction performance of different deep learning models are graphically depicted in Figure 3.

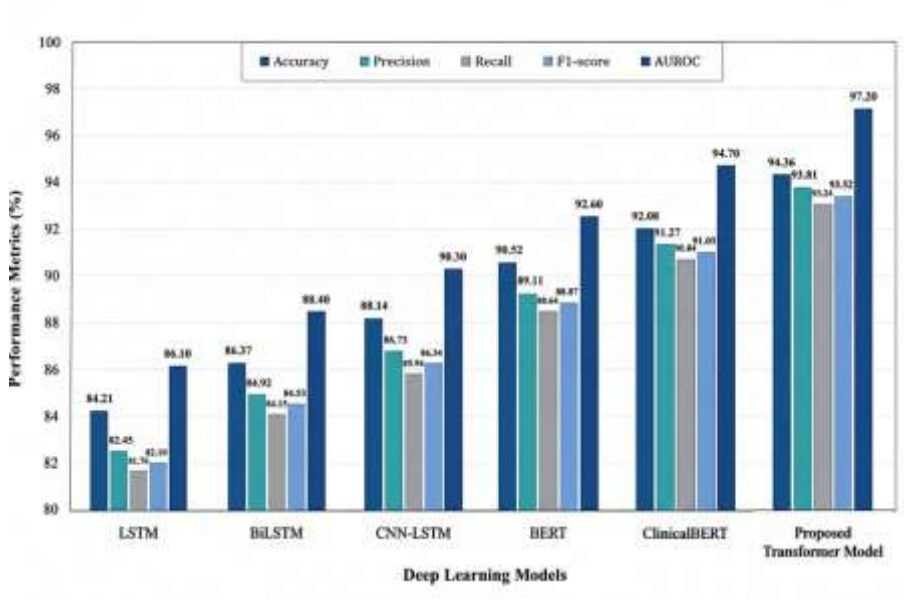

Fig 3. Diagnosis Prediction Performance Comparison Across Deep Learning Models.

The grouped bar chart shows the comparative analysis of accuracy, precision, recall, F1-score, and AUROC achieved by LSTM, BiLSTM, CNN-LSTM, BERT, ClinicalBERT and the proposed clinical NLP framework based on transformer. The proposed model consistently outperformed all the evaluation measures, showing that it has a better understanding of context and prediction capacity of clinical diagnosis.

4.2 EHR Summarization Performance Evaluation

The EHR summarization ability of the proposed clinical NLP framework based on transformer has been assessed and contrasted against multiple baselines such as TextRank, Seq2Seq, Pointer Generator, BART and ClinicalBERT Summarizer models. The assessment focused on the quality of the generated clinical summaries, their semantic preservation, understanding of the context, and their linguistic coherence. Comprehensive performance measures such as Standard healthcare NLP summarization metrics (ROUGE-1, ROUGE-2, ROUGE-L, BLEU score, BERTScore) were used to evaluate the performance. Table 2 shows the comparisons of the summarization results.

Table 2. Performance Comparison of EHR Summarization Models

Model	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	BLEU (%)	BERTScore (%)
TextRank	41.82	23.64	38.95	29.41	84.26
Seq2Seq	46.37	28.15	43.72	33.88	86.74
Pointer Generator	49.84	31.92	46.51	36.47	88.53
BART	53.28	35.76	50.84	40.92	91.14
ClinicalBERTSummarizer	56.73	39.41	54.22	43.68	93.27
Proposed Transformer Model	60.95	44.36	58.81	47.92	96.14

The significance of the difference is statistically significant, as indicated by Table 2 that shows the conventional extractive summarization methods like TextRank obtained relatively poor summarization performance since they lacked contextual understanding and were not able to produce clinically rich summaries. Moderate improvements were obtained in these models using sequential language learning mechanisms and attention-guided text generation mechanisms, which are the components of sequence-based deep learning models (such as Seq2Seq models or Pointer Generators). The presence of contextual embedding and improved semantic representation learning was one of the main factors that made architectures that employed transformer models, like BART or ClinicalBERT Summarizer, so effective for summarizing.

The proposed clinical NLP framework has the best performance on all evaluation metrics with ROUGE-1 score of 60.95%, ROUGE-2 of 44.36%, ROUGE-L of 58.81%, BLEU score of 47.92% and BERTScore of 96.14%. The results show a significant improvement, suggesting that the proposed model can accurately retain the clinically relevant information and include contextual relations in the longer medical narratives, as well as synthesise short and meaningful EHR summaries with high semantic meaning. Figure 4 shows how the quality of the summarization results obtained with different deep learning models differs from one another.

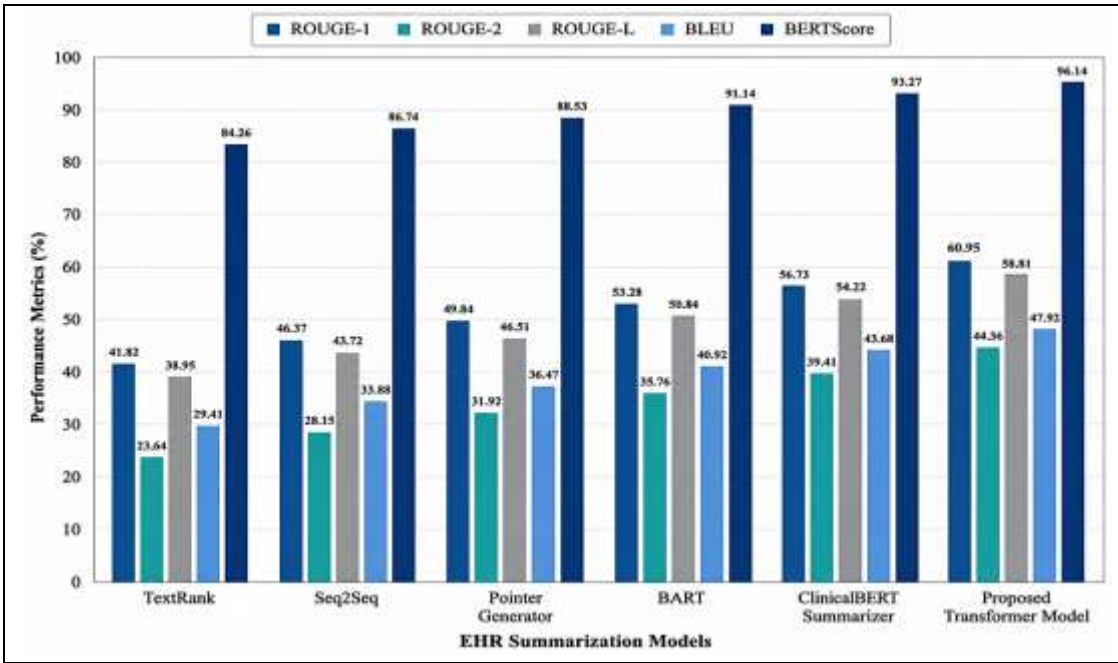

Fig. 4. Comparative Evaluation of EHR Summarization Quality Metrics Across Different Deep Learning Models.

Following is the grouped bar chart showing the comparative analysis in terms of ROUGE-1, ROUGE-2, ROUGE-L, BLEU and BERTScore of the summarized texts generated by TextRank, Seq2Seq, Pointer Generator, BART, ClinicalBERT Summarizer and proposed transformer-based clinical NLP framework. The proposed model achieved the highest summarization accuracy for all of the three metrics: semantic preservation, contextual understanding and clinical summarization ability, consistently.

4.3 Comparative and Ablation Analysis

To further validate the effectiveness of the proposed clinical NLP framework based on the transformer, comparative and ablation analysis was performed to see the contributions of different components and learning methods in architecture. The main goals of this analysis were to compare the proposed model to current healthcare NLP techniques and to investigate how attention mechanisms, multitask learning and pre-processing techniques affect the diagnosis prediction and summarization of EHRs performance.

The comparative evaluation showed that the transformer-based contextual representation learning method was clearly superior to the traditional recurrent and sequential deep learning methods. Classical models like LSTM and BiLSTM were found to be poor at modeling distant semantic information from complicated clinical text, resulting in poor diagnosis prediction accuracy and poor summarization. CNN-LSTMs achieved better extraction of local features, but performed comparatively weaker when it comes to contextual understanding than transformer-based approaches. While the models mainly focus on single-task settings in the healthcare NLP field, models like BERT and ClinicalBERT outperform with their self-attention learning and contextual embedding generation. The proposed framework further boosted performance by adding in multitask learning and shared semantic representation for simultaneous diagnosis prediction and EHR summarization.

To test the attention mechanism, we conducted an ablation study to assess how much it contributes to the transformer encoder architecture. The model exhibited significant performance declines when the multi-head self-attention module was either ablated or significantly reduced in size when tested on diagnosis prediction and summarization tasks. In particular, the model's ability to find clinically significant relationships between medical items and patient observations was decreased when contextual attention learning was omitted. The F1-score dropped by around 3-4% and ROUGE and BLEU scores also dropped due to poor semantic representation and the coherence generation in the context. This finding suggests attention guided contextual learning is an important factor in obtaining meaningful clinical information from unstructured narrative EHRs.

The effect of the multitask learning was also studied by comparing the proposed multitask framework with the independently trained diagnosis prediction and summarization models. The multi-task architecture benefited from the use of shared contextual feature representations which enhanced the semantic understanding of the shared data across both NLP tasks in the healthcare context. Shared learning helped to capture important medical concepts and prevented overfitting of redundant feature extraction processes to better generalize clinical patterns. The multi-task learning model significantly outperforms the single-task transformer models in terms of diagnosis prediction accuracy with around 2-3%, and also boosted summarization quality metrics like ROUGE-L and BERTScore on the basis of experimental observations.

Moreover, preprocessing methods were also important factors for achieving system performance. The clinical text normalization, tokenization, medical abbreviation expansion and ICD code mapping enhanced the semantic consistency and noise reduction in the healthcare datasets. The model struggled with convergence and limited contextual understanding of the medical terminology and noisy representations in text without preprocessing. On top of that, clinical entity extraction and standardized ICD mapping further boosted reliability of diagnosis classification by facilitating disease-related feature learning.

In conclusion, the comparative and ablation experiments validated the effectiveness of using the transformer-based attention mechanisms, multi-task learning and structured preprocessing operations to enhance the performance of EHR summarization and automated clinical diagnosis prediction in healthcare NLP systems.

5. Discussion

Experimental results prove that clinical NLP framework based on the proposed transformer model achieves significant improvements over the baseline models (conventional deep learning models) in terms of both automated diagnosis prediction and EHR summarisation. The main reason why the proposed model has performed well is due to its capability of contextual representation learning, multitask architecture, and attention-guided semantic feature extraction. Unlike conventional recurrent neural network models, the transformer encoder can capture long-range dependencies and context relationships in clinical narratives, which is particularly beneficial for interpreting unstructured clinical data to better recognize patients' condition and disease information.

The findings of the diagnosis prediction results proved that the proposed framework successfully provided a contextual understanding and also enhanced the disease classification ability in various clinical categories. Multi-head self-attention mechanisms facilitated the model's ability to detect relevant semantic relationships between symptoms, lab results, physicians' observations, and diagnostic patterns observed in electronic health records. This contextual learning proved to be very beneficial in terms of system’s predictive capability and the evaluation of classification performance such as accuracy, precision, recall, F1-score and AUROC. The proposed transformer framework showed superior performance in dealing with long clinical texts and minimizing information loss during feature extraction compared with the other models like LSTM, BiLSTM, and CNN-LSTM models. Moreover, the multitask learning strategy was found to be effective for improved generalization by sharing the semantic representations between the different tasks of the two subtasks. The improved clinical validation of the proposed model could prove valuable in supporting the diagnosis of a disease or in predicting the clinical risk of a patient.

The summarized results of the EHRs also show that the proposed framework can generate concise, meaningful and semantically coherent clinical summaries from long medical narratives. Existing extractive summarization methods usually lack contextual meaning and clinical relevance as they mainly depend on sentence selection instead of semantic understanding. By contrast, the proposed transformer-based summarization decoder employs contextual embeddings and attention-guided sequence generation to generate informative medical summaries, while filtering out redundant textual information. Enhanced ROUGE, BLEU and BERTScore confirms that the generated summaries are semantically close to reference clinical summaries. Computerized physician summary of concise EHRs can substantially help physicians with documentation tasks and make patient information more accessible in hospital settings.

The proposed framework is highly applicable to the deployment in intelligent healthcare systems and AI-enhanced clinical settings from a practical perspective. The framework can be used to build clinical decision support systems for doctors when they are dealing with prediction, monitoring and medical record analysis. Moreover, the integration with SIHS (smart hospital information systems) can enhance the automation of hospital processes and minimize paperwork-related administrative burdens. The model can also be used for telehealth platforms to help with the rapid interpretation of patients' records from a distance, enhancing access and efficiency in healthcare delivery. Additionally, the multitask transformer architecture offers a scalable structure for future advancements in medical AI for healthcare workflows, such as medical text mining, patient outcome prediction, and intelligent healthcare analytics.

Although the results are encouraging, the proposed study has a number of limitations. It can be seen that many clinical data have class imbalance issues, meaning that many disease categories have fewer number of training data, which could compromise prediction stability and generalization. Another challenge is domain adaptation as the style in which healthcare documents are presented differs between hospitals, for specific specialties, and geographical regions. The transformer-based architecture also calls for a considerable amount of computational power and graphics processing unit (GPU) memory for training and inference tasks with large datasets, potentially hindering its application in resource-limited healthcare environments. Before it's rolled out in real-world settings, privacy and ethical issues with patient data management, sharing EHR systems and the use of AI for clinical decisions need to be thoroughly addressed. Therefore, processing healthcare data securely, easily integrating AI explanations, and adhering to medical data protection regulations will be crucial for future deployments of healthcare NLP systems based on transformers.

6. Conclusion

This study introduced a clinical Natural Language Processing (NLP) framework based on a transformer model to predict the diagnosis automatically and extract the summary of the Electronic Health Record (EHR). The proposed architecture generated contextual embedding, extracted the feature with transformer encoder, learned the semantic information by the attention mechanism, and performed multitask representation learning for disease classification and clinical text summarization. When applied to unstructured healthcare sentences, the framework was able to effectively process them, ultimately yielding better predictive analysis results and semantic summary generation in a single NLP system for the healthcare domain.

The proposed framework was validated experimentally, showing its superiority over baseline models, including conventional deep learning models and existing transformer-based models. The diagnosis prediction module achieved better accuracy, precision, recall, F1 score, and AUROC results, showing better performance in the classification of clinically relevant disease patterns in the EHR. Likewise, the summarization module improved the ROUGE, BLEU, and BERTScore metrics, showing its ability to produce concise, semantically meaningful and context-aware clinical summaries. By combining transformer-based attention mechanisms with multitask learning, performance in both contextual understanding and semantic preservation was markedly enhanced, as was the overall understanding of healthcare texts.

The proposed framework also has a high practical value in the field of healthcare automation and intelligent medical information processing. By automating the diagnosis prediction process, clinicians can make more efficient decisions and better assess patient risks, whereas EHR summarization can alleviate the burden on physicians when it comes to documentation and make crucial patient information accessible and easily available. The framework can be applied to clinical decision support systems, smart hospital information systems, telemedicine systems and AI-assisted healthcare workflows.

In conclusion, the proposed framework for clinical NLP based on transformers holds promise for scalable and efficient clinical text analytics and intelligent information management in healthcare settings. The study emphasises the increasing potential of transformer architectures for the future of automated healthcare systems and the next-generation AI-based clinical decision support technologies.

7. Future Scope

The future potential of the proposed clinical NLP framework based on the transformer is high for further developments and the integration into real-life healthcare. Future studies are also needed to integrate Explainable Artificial Intelligence (XAI) methodologies to boost transparency and reportability of diagnosis prediction and EHR summarization results. Explaining the attention visualization, the feature attribution methods, and the reasoning mechanism with transformers, can help healthcare workers comprehend the rationale for automated predictions and generated summaries, which is essential for clinical decision making since it requires high reliability and trust.

A potentially new pathway is the use of federated healthcare NLP frameworks for privacy preserving distributed learning. Healthcare data is frequently sensitive and institution-specific, affecting the possibility of building a centralised model across various hospitals and medical centers. By allowing collaborative training of the transformer without patient data sharing, federated learning can boost model generalization while adhering to medical privacy laws, like HIPAA and GDPR. The distributed nature of the learning may also enhance the robustness of the learning for a variety of clinical environments and healthcare populations.

Future works can also involve implementing the proposed framework in real-time systems as deployed in hospital information systems and clinical decision support systems. The implementation of the model in the electronic healthcare system can aid in automated patient monitoring, quick diagnosis support, intelligent healthcare documentation, and workflow optimization in emergency and outpatient environments. Furthermore, the ability to optimize the inference in real-time and the light-weight architecture of the transformer could further enhance the feasibility of deployment in resource limited healthcare scenarios. Multilingual clinical text processing is also an interesting future research avenue. The current NLP systems in healthcare are mostly built to process English clinical notes, thus are not applicable in multi-lingual clinical record systems. Focusing on the area of global healthcare, medical task-oriented knowledge graph, and cross-lingual contextual embeddings, along with region-specific medical terminologies, can boost the global access to healthcare and intelligent handling of medical texts.

Moreover, the future adoption of NLP in the healthcare industry using transformers can include integration of sensor data from wearable devices, Internet of Things (IoT) medical devices as well as biomedical monitoring system in real-time. Using EHR narrative data in conjunction with physiological data like heart rate, blood glucose, electrocardiogram, and oxygen saturation levels can offer a deeper context for personalized healthcare analytics and clinical decision support for predictive outcomes. Well-integrated multimodal healthcare data could thus be used to create next-generation smart healthcare ecosystems with enhanced intelligent medical management and patient monitoring.

8. Ethical Considerations

Ethical, legal, and clinical issues in handling patient information and making medical decisions with the help of AI are significant considerations in the development and deployment of transformer NLP systems in healthcare. Patient privacy preservation is crucial in EHRs for data collection, preprocessing, model training, and deployment. All clinical datasets were de-identified with procedures to remove patient-related information and allow the secure handling of clinical data.

Responsible implementation of AI in healthcare requires compliance with data protection regulations, like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). The proposed framework aimed at facilitating privacy-preserving healthcare analytics was engineered to ensure minimal exposure of confidential patient information in the training and evaluation of models. Properly leveraging patient data in future real-world healthcare systems should also involve secure patient data storage, encrypted communication protocols, access control policies, and federated learning approaches to enhance patient data security and regulatory adherence.

Another ethical issue is Bias and Fairness in clinical AI systems. Biases stemming from uneven distribution of diseases, variations in demographics, and institutional-specific documentation styles can affect prediction

models built using transformer architectures like F-TCNN in the healthcare domain. This may have adverse impacts on the reliability of diagnosis and healthcare equity for various patient groups. To reduce these risks, balanced dataset preparation, data normalization, regularization techniques and cross domain evaluation techniques should be used during model development. In addition, future studies are needed to explore fair transformer-based architectures and explainable AI tools to enhance transparency and minimize the potential for unintended bias in automated clinical prediction.

While an increasing number of health care systems are being developed with the help of AI, safety and reliability in the clinical deployment of these systems still rely on human-in-the-loop validation. The proposed framework is successful in predicting the diagnosis and summarizing the EHR, but the summaries of the EHR should not be a substitute for medical judgment. Healthcare decisions should not be solely based on AI-generated diagnosis and summaries and should be reviewed by clinical experts. Incorporation of clinician feedback mechanisms and monitoring of the models in real-time can provide further improvement in system reliability, accountability, and patient safety in the real world.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78).
2. Devlin, J., Chang, M. W., Lee, K., &Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186).
3. Huang, K., Altosaar, J., &Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.
4. Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., &Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.
5. Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., ...& Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1.
6. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), 1–9.
7. Kormilitzin, A., Vaci, N., Liu, Q., &Nevado-Holgado, A. (2021). Med7: A transferable clinical natural language processing model for electronic health records. Artificial Intelligence in Medicine, 118, 102086.
8. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
9. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ...&Zettlemoyer, L. (2020, July). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880).
10. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.
11. Peng, Y., Yan, S., & Lu, Z. (2019, August). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 58–65).
12. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ...& Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
13. Rajkomar, A., Dean, J., &Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.
14. Shickel, B., Tighe, P. J., Bihorac, A., &Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ...&Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.