A Gaussian matrix graphical encoder in sports medicine diagnosis combining structured and unstructured data

Anonymous ACL submission

Abstract

We study the integration of Electronic Medical Records (EMRs) from clinical study into a joint predictive model. Compared to the totally black-box models, a competitive model with explainable structure is much more desirable. To tackle this challenge, this paper introduces a novel Gaussian Matrix Graphical Encoder(GMGE) based on matrix normal graphical model to encode unstructured medical text and simultaneously learn the underlying conditional dependency graph of concepts. We further present DiMES, a Diagnostic Model with Explainable Structure, which integrates the concept graph generated by GMGE with structured data such as patient's physical examination measures. Utilizing Graph Convolutional Networks (GCNs), DiMES encodes patient features based on the concept graph for downstream tasks, providing clinicians with accurate predictive information to assist in diagnostic decisions and treatment plan design. The effectiveness of the proposed DiMES is validated through its application on four downstream diagnostic predictive tasks(ACL, PCL, MMI and PS).

1 Introduction

The incorporation of Electronic Medical Record (EMR) data (including the outpatient records, MRI report and Physical Examination) into predictive analytics plays a pivotal role in the functionality of clinical decision-making systems. In recent years, the availability of large EMR data has enriched researchers with an abundant source of information and enabled deep learning methods for diverse tasks such as predictive diagnoses (Kopitar et al., 2020) and disease progression prediction(Zhang et al., 2019). Such predictions are vital for tailoring treatment plans, optimizing resource allocation, and improving patient outcomes.

Complex cases demand a profound comprehension of a patient's medical history as documented in electronic medical records, complemented by reasoning that incorporates medical images and various laboratory results. Consequently, numerous deep learning methods like (Wang et al., 2023) have emerged in an attempt to learn the internal structural relationships within medical record data. Recent advancements in the field of natural language processing and machine learning have brought to the forefront the Transformer architecture, originally introduced by (Vaswani et al., 2017). The potential of transformer has been noticed in the medical sector, where pioneering studies such as BEHRT by (Li et al., 2020b) and Transformehr by (Yang et al., 2023) have begun to explore its applications. These models provide a sophisticated methodology for managing the complex characteristics of EMR data. They show their considerable potential to significantly augment medical predictive analytics. However, clinical professionals demand models that are both predictive and explainable. Transformer models

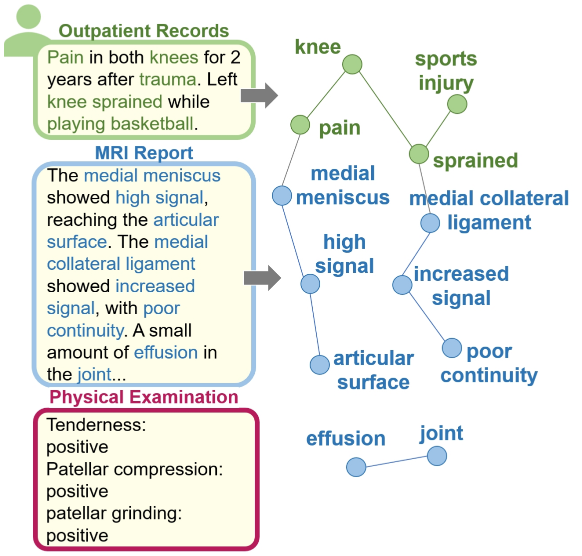

Figure 1: Medical Records with Explainable Graph Structure

often lack the necessary explainability, making it difficult for clinicians to understand and trust the results, which is critical for their practical application.

An effective method to enhance model interpretability is through the use of graph structures. Leveraging graph structures allows experts to gain a clearer understanding of the model's reasoning logic and to verify its clinical interpretability. (Choi et al., 2020) proposed the Graph Convolutional Transformer (GCT) for learning EHRs' hidden encounter structures. The model disregards the relationships between nodes within the same category, which is essential for symptom-based predictive diagnostics, as treatment information is frequently not available in predictive diagnostics tasks. MedPath(Ye et al., 2021) extracts personalized knowledge graphs (PKGs) from large-scale online medical knowledge bases. It utilizes a graph neural network encoder to learn embeddings of the PKGs, thereby achieving enhanced predictive capabilities. (Wu et al., 2023) constructs an EHR hypergraph and employs a multi-view learning framework, the method is capable of capturing higher-order relationships between patient visits and medical codes.Other similar methods (Liu et al., 2020), (Meng et al., 2021) have shown that graph structure are a good tool for improving performance and interpretability. Utilizing graph structures for modeling also allows for a more effective integration of prior medical external knowledge into the model (Ma et al., 2018), (Ye et al., 2021) , endowing the model with medical reasoning abilities based on established medical theories. Explainable mechanisms such as knowledge graph modeling can significantly increase model complexity and often require extensive annotated prior datasets. We need a simpler method capable of unsupervised learning the graph structure of concepts from medical record data, especially from unstructured text data. (Lai and Yin, 2024) attempted to learn conditional dependence graph using GloVe embeddings, but this approach assumes that the word embeddings must follow a matrix normal distribution. In contrast, we directly encode or transform concepts representations into a matrix normal distribution.

We propose a straightforward and intuitive method for integrating medical knowledge, a model that simultaneously encodes unstructured text and generates graph structures in an unsupervised manner, displaying the relationships between key concepts in patient medical records, as shown in the figure1.

We have designed a novel encoder to encode unstructured text, generating a corresponding conditional dependency graph of concepts during the encoding process based on Matrix Normal Graphical Model (MNGM). This allows us to conveniently incorporate prior information into the constraints of the concept graph. Our approach does not require extensive training or complex priori knowledge bases, nor does it significantly increase model complexity. It is capable of effectively capturing and clearly presenting the structural relationships within patients' electronic medical record data.

Contributions:

- We design an Gaussian Matrix Graphical Encoder (GMGE) to encode unstructured text and learn the underlying graph structure of concepts simultaneously. It utilizes the penalized likelihood function of the matrix normal graphical model, learning the precision matrix between medical concept representations, thereby constructing a graph of conditional dependencies between concepts.

- Guided by medical experts, we have established a series of regular expression rules to extract key concepts from unstructured Chinese medical texts such as patient complaints and MRI reports. Each concept is integral to the diagnostic process, providing key information that aids in identifying and understanding a patient's condition.

- We propose a Diagnostic Model with Explainable Structure (DiMES) for multiple diagnostic prediction tasks. The model encode structured and unstructured data separately, using GMGE to encode medical texts and output key concept graphs. The graph structure enhances model explainability, fostering trust among clinicians and patients.

The model proposed in this study is designed to provide clinicians with more accurate predictive information to assist them in making diagnostic decisions and designing treatment plans.

2 Models

2.1 MNGM

Assume the data Y as a matrix-valued random variable, we say Y follows a matrix normal distribution, if Y has a density function

$$p(Y|M,U,V) = k(U,V)\exp(-\mathrm{tr}\{(Y-M)^{\top} U^{-1}(Y-M)V^{-1}/2\}), \quad (1)$$

where $k(U,V) = (2\pi)^{-pq/2}|U|^{-q/2}|V|^{-p/2}$ is the normalizing constant, M is the mean matrix, U is the row-covariance matrix and V is column covariance matrix. This definition is equivalent to the definition via the Kronecker product, specifically,

$$Y \sim MN_{p,q}(M;U,V) \quad \text{if and only if}$$

$$\mathrm{vec}(Y) \sim N_{pq}(\mathrm{vec}(M), V \otimes U).$$

We denote the corresponding precision matrices as $A = U^{-1}$, $B = V^{-1}$ for U and V, respectively. This model assumes a particular decomposable covariance matrix for $\mathrm{vec}(Y)$ that is separable in the geostatistics context (Cressie, 1993).

The following proposition shows that there is a graphical model interpretation for the two precision matrices A and B in the matrix normal model (1). See reference in (Yin and Li, 2012).

Proposition 1 *Assume that $Y \sim MN_{p,q}(M;U,V)$. If we partition the columns of Y as $Y = (Y_1, \cdots, Y_q)$, then it holds for $\gamma, \mu \in \Gamma = \{1, \cdots, q\}$ with $\gamma \neq \mu$ that*

$$Y_\gamma \perp\!\!\!\perp Y_\mu \mid Y_{\Gamma\setminus\{\gamma,\mu\}} \text{ if and only if } b_{\gamma\mu} = 0,$$

where $B = \{b_{\alpha\beta}\}_{\alpha,\beta\in\Gamma} = V^{-1}$ is the column precision matrix of the distribution; similarly, if we partition the rows of Y as $Y = (Y^1, \cdots, Y^p)^T$, then it holds for $\delta, \eta \in \Delta = \{1, \ldots, p\}$ with $\delta \neq \eta$ that

$$Y^\delta \perp\!\!\!\perp Y^\eta \mid Y^{\Delta\setminus\{\delta,\eta\}} \text{ if and only if } a_{\delta\eta} = 0$$

where $A = \{a_{\delta\eta}\}_{\delta,\eta\in\Delta} = U^{-1}$ is the row precision matrix of the distribution.

We estimate the precision matrices $A = U^{-1}, B = V^{-1}$ in model (1) by a penalized likelihood estimation. To estimate the A and B, one can minimize the following penalized negative log-likelihood function

$$\begin{aligned}\phi(A,B) &= -q\log(|A|) - p\log(|B|) \quad (2)\\ &\quad + \frac{1}{n}\sum_{k=1}^{n}\mathrm{tr}\{AY_kBY_k^T\}\\ &\quad + \sum_{i\neq j} p_{\lambda_{ij}}(a_{ij}) + \sum_{i\neq j} p_{\rho_{ij}}(b_{ij})\end{aligned}$$

where $p_{\lambda_{ij}}(\cdot)$ is the penalty function for the element a_{ij} of A with tuning parameter λ_{ij}, while $p_{\rho_{ij}}(\cdot)$ is the correksponding penalty function for b_{ij} with tuning parameter ρ_{ij}.Here we use lasso penalty function $|\cdot|_1$ as $p_{\lambda_{ij}}(\cdot)$ and $p_{\rho_{ij}}(\cdot)$. We tune the penalty parameters λ_{ij} and ρ_{ij} by controlling the output amount of edges on the graph at certain level.

2.2 GMGE

Gaussian Matrix Graphical Encoder (GMGE) is designed to encode unstructured medical text and learn the underlying conditional dependence relationships among concepts embedded in a semantic space. By encoding concepts into Matrix Normal distributions, leveraging principles from graphical models and Gaussian distributions, GMGE provides a robust framework for the hierarchical encoding process.

Assume Y as the representation of concepts. From , we derive a negative penalized likelihood function when $n = 1$ and $B = I$:

$$\begin{aligned}P(A,Y) &= -q\log(|A|) + \mathrm{tr}\{AYY^T\}\\ &\quad + \sum_{i\neq j} p_{\lambda_{ij}}(a_{ij}) \quad (3)\end{aligned}$$

Based on this likelihood function3, representations of concepts can be transformed into a matrix normal distribution characterized by a sparse precision matrix. The loss function of the GMGE is designed as:

$$L = L_{MLM}(Y) + \omega P(A,Y)$$

L_{MLM} means the loss function of the Masked Language Model task from BERT.(Devlin et al., 2018) ω is a weighting parameter. Pre-trained embeddings of concepts, noted as $M_{p\times q}$ are just one implementation in the semantic space. So we assume the underlying concepts embeddings variables in the semantic space follows a Matrix Normal distribution denoted as $Y \sim MN(M,U,I)$. U represents the covariance matrix of concepts. $A = U^{-1}$ stands for the row-precision matrix. Additionally, we make the assumption that the dimensions of word embeddings are independent, thus leading to an identity covariance matrix denoted by I. The $M^{(k)}$ matrix in kth batch will be $Y^{(k-1)}$. This means in every batch we learn the representation $Y^{(k)}$ based on the mean of last batch version of embeddings, thus update the Y and A step by step. We use L_{MLM} to update

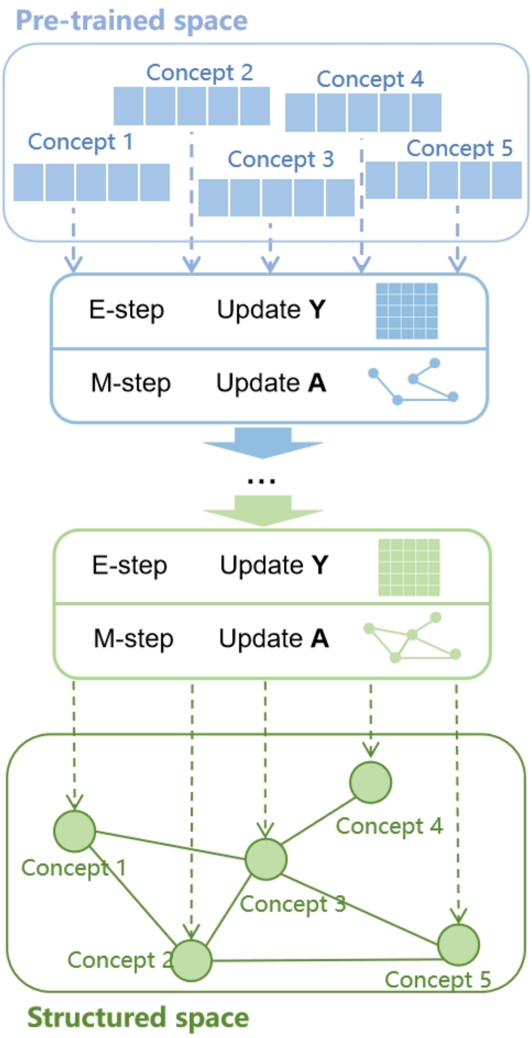

Figure 2: Framework of GMGE: Pre-trained representation will be iteratively transferred into a structured space while simultaneously updating emceddings matrix Y and the graph structure (represented in the form of an accuracy matrix).

Y, and then update precision matrix A based on likelihood P, thus update the Y and A step by step. This process is similar to EM algorithm.

If we consider the MLM (Masked Language Model) loss task as seeking word embeddings that approximate the true distribution of the masked word, then maximizing the negative MLM loss is equivalent to performing likelihood estimation on Y with the true distribution function, in order to maximize the likelihood function.

E-step (Expectation step): In the EM algorithm, the E-step is tasked with calculating the expected values of the latent variables based on the current model parameters. In the context of the GMGE algorithm, using the precision matrix A and the mean matrix M from the previous step as parameters, we calculate the expected value of Y under the likelihood distribution function $-L$.

M-step (Maximization step): The M-step of the EM algorithm focuses on maximizing the expected function to update the model parameters. For the GMGE algorithm, After updating Y, we substitute it back into the function and update A by maximizing the likelihood function.

The EM algorithm iteratively updates the model parameters by alternating between the E-step and M-step. Similarly, the GMGE algorithm iteratively optimizes model parameters, while incorporating the pre-training tasks of BERT and the properties of matrix normal distributions, thereby facilitating effective learning of conceptual representations.

Upon completion of learning, we obtain the the conditional dependence graph corresponding to precision matrix A and the updated concepts representations. We further encode concepts using GCNs based on downstream tasks. This fine-tuning step allows us to integrate the relationships and attributes of the medical concepts into a comprehensive representation.

2.3 Diagnostic Model

The EMRs are divided into two main components: unstructured data like outpatient record texts and MRI reports; structured data like physical examination results. The unstructured texts primarily consist of patients' chief complaints, present illness histories, and MRI reports recorded by outpatient physicians and radiologists. The structured data are mainly discrete data reflecting the assessments made by physicians during the physical examination of specific items.

We acquire a key concepts list based on the frequency of concepts occurrence, regular expression rules, and professional guidance from physicians. The types of concepts include categories such as body parts, structures, pathologies, symptoms, severities, etiologies, and treatments. Utilizing these concepts as nodes, we employ the GMGE method to derive a concepts graph and node embeddings. By utilizing the concepts present in each patient's medical record text, we obtain node embeddings for each patient involved and pool them to form the patient's features. These features are then concatenated with structured data information to derive the patient's final feature. For each patient, we can generate a relationship graph of the key concepts contained, thereby obtaining an intuitive explanatory graph for diagnostic

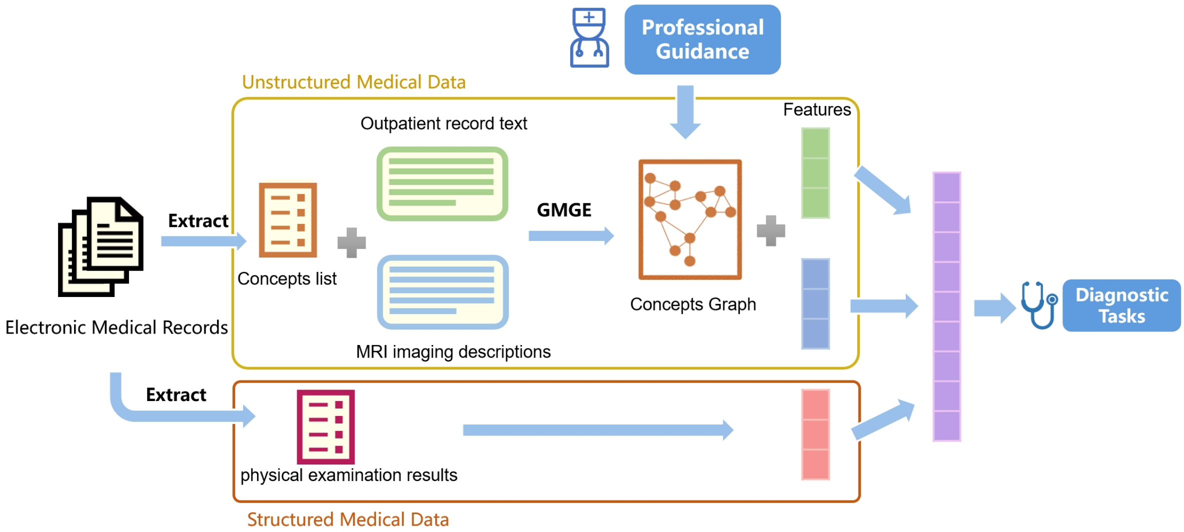

Figure 3: Framework of Diagnostic Model: The model separately extracts structured and unstructured data from EMRs. The unstructured text is processed by GMGE, which encodes the text and produces a graph structure. The graph is then utilized by a GCN to generate the final patient features. The structured data is encoded into one-hot vectors and concatenated with the features extracted from the text to perform predictive analytics tasks.

predictions. We further encode patients features using Graph Convolutional Networks (GCNs) based on the concepts graph and downstream tasks. We note this Diagnostic Model with Explainable Structure as DiMES.

The complete framework is shown in Figure 3.

3 Experiment

3.1 Dataset

We possess electronic medical records in Chinese from 3,399 patients diagnosed with knee joint conditions at the Sports Medicine Department of the Peking University Third Hospital.Each patient's medical record in this consists of three distinct parts:

- Outpatient Records: The outpatient physician's records of the patient's chief complaints and the history of the present illness, which are text data that every patient have.
- MRI Reports: The report written by the radiologist based on the patient's MRI images, which are text data that do not include the specific MRI diagnoses. For patients who have not undergone MRI scans, this section would be absent in the medical record. (1,568 patients have MRI reports in our dataset.)
- Physical examination results: records of physical examinations performed by outpatient doctors on patients, including the results of 35 items such as patellar grinding and patellar compression tests, which are discrete structured data that every patient have.

Moreover, we have identified 240 diagnostic keyword concepts derived based on professional guidance, regular expression rules, and the frequency of occurrence of terms within the medical records. They are all concepts that play a key role in knee joint diagnosis and appear frequently in medical records. The keyword concepts include: Boat wedge joint, joint cleaning, medial knee, quadriceps femoris, sprains, cycling, free combat, etc.

3.2 Experiment Settings

We used a fine-tuned BERT model from (Li et al., 2020a), which was pre-trained on Chinese clinical corpora, as our initial model. Then we continue train the model on the outpatient record text and MRI imagine descriptions in medical records using GMGE. This process was performed on one NVIDIA GeForce RTX 3090 for 16 hours. After this process, we obtain embeddings for each key concept and the precision matrix A that illustrates how these concepts are interconnected.We select the strongest relationships to form a concept relationship graph by examining the absolute values of the elements in A. We then apply a Graph Convolutional Network to fine-tune the model for downstream tasks. The concepts present in a patient's medical record form a subgraph of

Model	ACL	PCL	MMI	PS	Average
Bert	88.93%	96.73%	85.37%	66.61%	84.41%
GMGE with random graph(50)	89.64%	**97.50%**	87.61%	72.20%	86.73%
DiMES w/o guidance (50)	92.03%	97.49%	**88.70%**	72.37%	**87.64%**
DiMES w/o guidance (100)	91.76%	**97.50%**	87.85%	72.05%	87.29%
DiMES (50)	**92.17%**	**97.50%**	88.11%	72.23%	87.50%
DiMES (100)	91.56%	**97.50%**	87.70%	**72.50%**	87.32%

Table 1: Results of different models, using the average accuracy as criterion. The number after the model name indicates the number of edges in the concepts graph associated with that model.

the concepts graph. We perform pooling on the nodes contained within this subgraph and then concatenate the resulting features with physical examination features to predict the patient's diagnosis. The specific predictive tasks include:

- Predicting abnormalities in the anterior cruciate ligament (ACL);
- Predicting abnormalities in the posterior cruciate ligament (PCL);
- Predicting medial meniscus injuries(MMI);
- Predicting patellar softening(PS).

The fine-tuning process was performed on NVIDIA GeForce RTX 3090 for about 20 minutes per task.

3.3 Results

We compare the results among Bert, Bert with random graph, GMGE without professional guidance, and GMGE with professional guidance. Professional guidance would mark 10 pairs of concepts as connected by edges and 10 pairs of concepts as not connected by edges, leaving the remaining pairs unaltered without any intervention. We employ a five-fold cross-validation to obtain our average accuracy results. The results are shown in Table 1

We also compared the performance of our model under different conditions to ensure comprehensive data mining from each segment of information: using only outpatient record texts, using a combination of outpatient records and MRI report texts, and using outpatient records, MRI report texts, along with physical examination information, ensuring that the model is effectively leveraging information from all parts. The results of predicting abnormalities in the anterior cruciate ligament are shown in Table 2. The node edges output by our model, such as "Cold stimulation - Patella," suggest a potential etiology for issues in the patellar region; "Cruciate ligament - Continuity" indicates the MRI imaging characteristics of the cruciate ligament; "Knee - Crepitus" represents a possible clinical manifestation in the knee area, and so on. These outputs illustrate how the model can capture and represent the relationships between different medical concepts, providing a structured way to understand the complex interactions within patient data. In Figure 4, we present an example of our output concepts relationship in a certain patient. This visualization demonstrates the reasoning logic our model takes into account when making predictions.

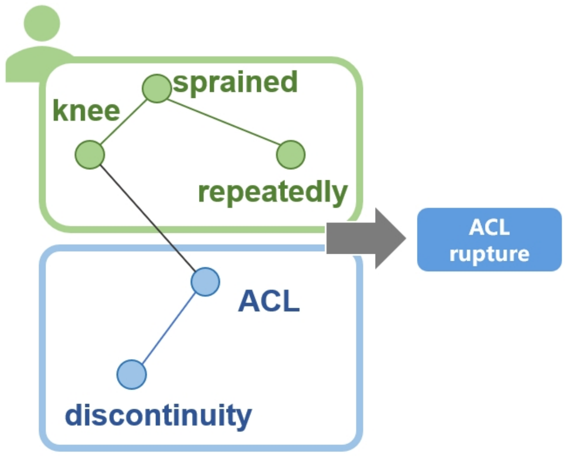

Figure 4: An example of a patient's EMR structure output from DiMES

4 Conclusion

In this paper, we propose GMGE, an encoder based on the matrix normal graphical model, and further build a graph-interpretable diagnostic model DiMES for multiple diagnostic prediction tasks. GMGE's loss function is derived from two components: the loss from the Masked Language Modeling (MLM) task and the penalized likelihood

DiMES(50)	Accuracy	Precision	Recall
Outpatient records	91.14%	82.42%	71.43
Outpatient records + MRI reports	92.00%	83.23%	72.30
Outpatient records + MRI reports + Physical examination results	92.17%	85.40%	71.69%

Table 2: Results of DiMES using different parts of data

function of the matrix normal distribution. We iteratively update the concept embeddings and the graph structure corresponding to the precision matrix using an EM-type algorithm, obtaining the conditional dependency graph between concepts. We then combine the concept relationship graph with a GCN to predict the different diagnostic outcomes of patient medical records, resulting in a multi-task diagnostic model.

5 Limitation

The encoder in this paper can be improved to use the LLM with world model pre-trained in it. Also, we consider only text corpus data and not including the medical imaging data maybe a significant drawback of our method. While the prediction results alleviate this concern by noting that the accuracy is quite acceptable in medical practice. The image data analysis should be carefully modeled in a multi-modal framework, which is beyond this paper's scope.

References

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. 34(01):606–613.

Noel AC Cressie. 1993. *Statistics for spatial data*. John Wiley & Sons, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1):11981.

Jizheng Lai and Jianxin Yin. 2024. Learning conditional dependence graph for concepts via matrix normal graphical model. *Statistics and Its Interface*, 17(2):187–198.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020a. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of Biomedical Informatics*, 107:103422.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020b. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155.

Zheng Liu, Xiaohan Li, Hao Peng, Lifang He, and S Yu Philip. 2020. Heterogeneous similarity graph neural network on electronic health records. In *2020 IEEE international conference on big data (big data)*, pages 1196–1205. IEEE.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 743–752, New York, NY, USA. Association for Computing Machinery.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. 2021. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25(8):3121–3129.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2852, Singapore. Association for Computational Linguistics.

Jialun Wu, Kai He, Rui Mao, Chen Li, and Erik Cambria. 2023. Megacare: Knowledge-guided multi-view hypergraph predictive framework for healthcare. *Information Fusion*, 100:101939.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. 2023. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. pages 1397–1409.

Jianxin Yin and Hongzhe Li. 2012. Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis*, 107:119–140.

Yuan Zhang, Xi Yang, Julie Ivy, and Min Chi. 2019. Attain: Attention-based time-aware lstm networks for disease progression modeling. pages 4369–4375.