Multi-Length Meta-Path Sematic Fusion in Medical Heterogeneous Graph for Disease Dignosis

Jianbin Luo, Dan Yang, Yang Liu

Abstract—**Heterogeneous graph neural networks have attracted significant attention in the field of disease diagnosis. Medical heterogeneous graphs encompass various types of nodes and edges, representing rich medical information and interconnections. However, there are limitations in applying inherited attention and multi-layer structures from graph neural networks to disease diagnosis tasks. Firstly, introducing attention to large medical heterogeneous graphs leads to significant computational complexity. Secondly, employing multi-layer structures when dealing with large medical heterogeneous graphs, with each layer performing semantic fusion, may cause semantic confusion and easily lead to issues such as vanishing or exploding gradients. To address these issues, a multi-length meta-path sematic fusion in medical heterogeneous graph for disease dignosis (MLM4DD) has been proposed. MLM4DD uses a lightweight average aggregator to precompute neighborhood aggregation, reducing computational complexity and improving information propagation efficiency. To better utilize semantic information and avoid issues like vanishing and exploding gradients, MLM4DD introduces a single-layer structure with multi-length meta-paths to expand the receptive field. It incorporates local attention and multi-scale attention fusion to capture features from different meta-paths, thus obtaining embedded representations of patient nodes. Extensive experiments on the MIMIC-IV dataset demonstrate that MLM4DD outperforms other baseline methods in terms of disease diagnostic performance, effectively improving the accuracy of disease diagnosis.**

Index Terms—**Disease Diagnosis, Electronic Medical Records, Multi-Length Meta-Path, Medical Heterogeneous Graph**

I. INTRODUCTION

With the accumulation of medical big data, Electronic Medical Records (EMR) have shown significant potential in personalized healthcare services, particularly in disease diagnosis [1-2] and similarity [3]. EMR refers to detailed records of clinical events during a patient's visits and hospitalizations, including patient demographics, medication treatments, medical procedures (such as dual catheter coronary angiography, single-vessel surgery, vascular bifurcation surgery, etc.), diagnostic information, and various laboratory test results. It represents a comprehensive collection of diverse medical data. This study primarily focuses on disease diagnosis based on electronic medical records, aiming to identify potential diseases that patients may have based on the information recorded in their electronic medical records.

Recently, Graph Neural Networks (GNN) have achieved notable progress. GNN primarily targets homogeneous graphs, where nodes and edges share a single type, using neighborhood aggregation strategies to capture structural information. This method allows GNN to effectively learn relationships between nodes and the overall graph structure. However, GNN show certain limitations when addressing heterogeneous graphs rich in semantic information. For instance, medical heterogeneous graphs include three types of nodes: patients (P), drugs (D), and procedures (O), along with two types of edges: patient-drug (indicating patients taking a specific drug) and patient-procedure (indicating patients undergoing a specific procedure). This heterogeneous graph structure not only involves complex associative relationships but also encompasses multi-level semantic information.

To overcome this challenge, various Heterogeneous Graph Neural Networks (HGNN) have emerged. These models focus on capturing semantic information within heterogeneous graphs and exhibit strong performance in heterogeneous graph representation learning. However, existing HGNN inherit many mechanisms from Graph Neural Networks, particularly attention mechanisms and multi-layer structures. In large medical heterogeneous graphs, attention increases computational complexity. This increase leads to decreased information propagation efficiency and reduced accuracy in disease diagnosis. Applying multi-layer structures to large medical heterogeneous graphs may cause difficulties in distinguishing high-level semantics. It can also result in issues such as vanishing or exploding gradients, significantly impacting disease diagnosis effectiveness. Related work [4] categorizes attention into two types: neighbor attention among neighbors within the same relationship and semantic attention between different relationships. It confirms that semantic attention is necessary, whereas neighbor attention is not. A single-layer structure with long meta-paths proves superior to a multi-layer structure with short meta-paths. As shown in Fig. 1(a), the medical heterogeneous graph composed of

Manuscript received Dec 28, 2023; revised Sep 17 2024.This work was supported by the General Scientific Research Project from the Educational Department of Liaoning Province (LJKMZ20220646).

Jianbin Luo is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: 17641241848@163.com).

Dan Yang is a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (corresponding author to provide e-mail: asyangdan@163. com).

Yang Liu is an associate professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: liuyang_lnas@163.com).

Electronic Medical Records (EMR) data includes three node types: patients (P), drugs (D), and procedures (O). It comprises two edge types: patient-drug (indicating patients taking a certain drug) and patient-procedure (indicating patients undergoing a specific procedure). Meta-paths are widely used structures for capturing semantics within heterogeneous graphs. Fig. 1(b) illustrates short meta-paths (l-hop <= 2): patient-drug (PD), patient-drug-patient (PDP), and patient-procedure-patient (POP). The PD meta-path details the relationship between patients and drugs. The PDP meta-path shows relationships between patients taking the same drug, while the POP meta-path describes relationships between patients undergoing the same procedure. Long meta-paths (l-hop > 2) include patient-procedure-patient-drug (POPD), patient-drug-patient-drug-patient (PDPDP), and patient-procedure-patient-procedure-patient (POPOP). The POPD meta-path describes patients and drugs associated with procedures. The PDPDP meta-path illustrates similar patients due to drug-related connections, while the POPOP meta-path highlights similar patients due to procedure-related links. Increasing the maximum meta-path length yields meta-paths with different semantics.

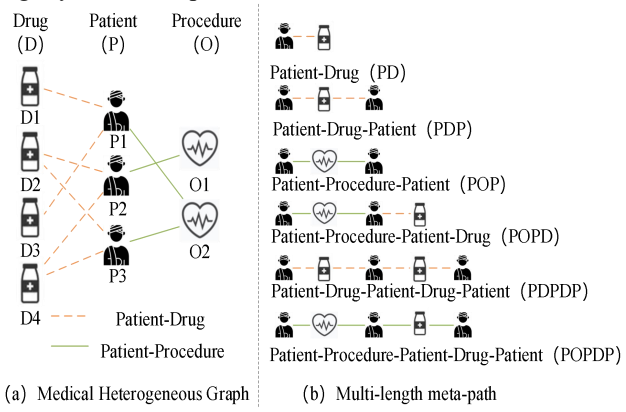

Fig.1. An example of medical heterogeneous graph and multi-length meta-path

To address the aforementioned challenges, propose a multi-length meta-path semantic fusion framework for disease diagnosis in medical heterogeneous graphs, named MLM4DD. The framework first utilizes electronic medical records to construct a medical heterogeneous graph. Subsequently, it employs an average aggregator to simplify neighborhood aggregation [5]. By eliminating redundant neighbor attention and avoiding unnecessary neighbor aggregations at each training stage, this approach reduces complexity and enhances information propagation efficiency. The framework adopts a single-layer structure with multiple-length meta-path to capture more distant relationships between nodes, integrating global information from the medical heterogeneous graph. This effectively captures semantic information within the heterogeneous graph and mitigates issues such as gradient vanishing and exploding. Finally, local attention and multi-scale attention are introduced to fuse semantics from different meta-paths, providing a more comprehensive embedding for patients. The design of this framework aims to improve the accuracy and generalization capability of disease diagnosis.

The primary contributions can be outlined as follows:

- Propose a multi-length meta-path semantic fusion framework for disease diagnosis in heterogeneous medical graphs. Fully utilize medical data in electronic health records to construct a heterogeneous medical graph. Use an average aggregator to simplify neighbor aggregation, reduce complexity, and improve information transmission efficiency. Employ a single-layer structure based on multi-length meta-paths instead of a multi-layer structure with short meta-paths to avoid issues such as gradient vanishing and gradient explosion.
- In the framework, robustness improves through channel-shuffling convolutional layers. These layers project semantic vectors from different meta-paths into the same feature space. Local attention and multi-scale attention fuse features from various meta-paths. This process provides a more comprehensive embedding for patients.
- Extensive experiments on the MIMIC-IV dataset to validate the feasibility and effectiveness of the MLM4DD framework. The experimental results show that the MLM4DD framework outperforms other mainstream frameworks.

II. Related Work

This section discusses work related to graph neural network disease diagnosis, including heterogeneous graph neural networks and disease diagnosis based on heterogeneous graph neural networks.

A. Heterogeneous Graph Neural Networks

In recent times, numerous heterogeneous graph neural network (HGNN) models have emerged. The design of HGNN models primarily focuses on modeling heterogeneous information. HGNN can be broadly classified into two categories. The first category is based on meta-path methods, as seen in related works [6-9]. These models initially capture structural information with similar semantics using various meta-paths and then integrate diverse semantic information. Models in this category aggregate neighborhood features within each meta-path's scope, generate semantic vectors, and subsequently fuse them to form the final embedding vector. The second category involves methods without meta-paths, as seen in related works [10-13]. These models simultaneously capture structural and semantic information. They aggregate local neighborhood messages (such as Graph Neural Networks or GNN) and embed semantic information into the propagated messages using additional modules like attention mechanisms. Simple-HGN [14] introduces a straightforward yet powerful GAT-based baseline model. This model simultaneously considers edge type embeddings and node embeddings to compute attention scores. HetGNN [15] defines semantic relationships between different types of nodes using meta-paths and aggregates features of different node types through Bi-LSTM. HAN [16] proposes a heterogeneous graph attention network that employs meta-paths to capture semantic relationships between various types of nodes and combines hierarchical attention mechanisms to learn both node-level and semantic-level structures. However, it remains constrained by the limitations of short meta-paths and multiple layer structures.

MAGNN [17] further utilizes all nodes in meta-path instances, not just those at the two endpoints.

B. Disease Diagnosis Based on Heterogeneous Graph Neural Networks

In electronic medical records, modeling complex objects and their various relationships is necessary. Combining heterogeneous graph neural networks with disease diagnosis can assist the healthcare domain in leveraging diverse types of medical data more accurately. This approach improves disease diagnostic and predictive performance and provides more personalized medical services for patients. In related work [18], the healthcare graph convolutional network (HealGCN) based on electronic health records is introduced. It employs graph convolutional networks to serve new users while using a symptom retrieval system to address the sparsity of medical description data. In related work [19], random resampling balances the dataset; graph convolutional neural networks (GCN) extract global features, and bidirectional self-attention networks (BERT) are integrated. The proposed VGBNet model aims to fuse local and global features for disease diagnosis and prediction. In related work [20], an adaptive graph learning method capable of automatically capturing latent graph structures is introduced. Based on this method, an end-to-end multimodal graph learning framework (MMGL) is proposed for multimodal disease prediction tasks.

The key difference between the proposed medical heterogeneous graph disease diagnosis framework with multi-length meta-path semantic fusion in this paper and the previous research lies in:

- Current models for disease diagnosis that utilize heterogeneous graph neural networks overlook the complexity and reduction in information propagation efficiency caused by attention mechanisms. In contrast, the proposed framework in this paper employs an average aggregator to simplify neighbor aggregation. This approach reduces complexity and enhances information propagation efficiency. Consequently, it learns more accurate patient embeddings.
- Existing heterogeneous graph neural networks often adopt a strategy of short meta-paths and a multi-layer structure to obtain rich semantic information. However, they overlook the challenges of applying a multi-layer structure to large-scale medical heterogeneous graphs. Executing semantic fusion at each layer leads to higher-level semantic confusion and is prone to issues like gradient vanishing or exploding. The proposed framework in this paper utilizes a single-layer structure based on multi-length meta-path. This approach captures more distant relationships between nodes, integrates global information in the medical heterogeneous graph, effectively captures semantic information, and avoids problems related to gradient vanishing and exploding.
- The current trend in integrating information from different meta-paths often involves the use of self-attention or multi-head attention mechanisms. In contrast, this paper adopts a combined approach of local attention and multi-scale attention. This approach focuses on both the local structure of nodes and information at various scales, resulting in a more comprehensive embedding of patients.

III. Preliminaries

The disease diagnosis framework defines key concepts as follows:

Definition 1. Medical heterogeneous graph. The medical heterogeneous graph is defined as $G = \{V, E, A, R\}$, where V represents the set of all nodes, and E represents the set of all edges. It is associated with the node type mapping function ϕ and edge type mapping function Ψ. Each node $v \in V$ has a mapping relation $v \to \phi(v)$, and each edge $e \in E$ has a mapping relation $e \to \Psi(e)$. A and R represent sets of node types and edge types, and $|A| + |R| > 2$.

Definition 2. Meta-path. The meta-path M defines composite relationships involving several edge types. Its form is $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$ (abbreviated as $A_1 A_2 \ldots A_{l+1}$). It describes a composite relationship between node types A_1 and $A_{l+1}: R = R_1 \circ R_2 \circ \cdots \circ R_l$, where $\circ$ represents the composite operator on the relationship. Given an instance of a meta-path $M(A_1, A_{l+1}) = \{A_1, R_1, A_2, R_2, \ldots R_l, A_{l+1} : A_i \in A, R_i \in R\}$. Specifically, $M(A_1, A_{l+1})$ represents the relationships in the l-hop neighborhood, where A_1 is the target node, and A_{l+1} is one of the neighborhoods based on the meta-path A_1.

The commonly used symbols in this paper are specified as shown in Table I.

TABLE I
SYMBOLS AND THEIR MEANINGS

Symbol	Meaning
G	Medical heterogeneous graph
V,E	Node and edge sets
A,R	Node type and edge type sets
X^c	Raw feature matrix
Y	Raw label matrix
Φ	Meta-path collection
M	Meta-path
h_i	Patient embedding
$\hat{y}$	Disease diagnosis results

IV. Disease Diagnosis Framework

The proposed disease diagnosis framework MLM4DD is illustrated in Fig.2. Firstly, the medical heterogeneous graph uses healthcare data from electronic medical records. Subsequently, multiple-length meta-paths aggregate using simplified neighborhood aggregation to obtain patient embeddings from each meta-path. Next, patient embeddings project into the same feature space using feature projection with channel shuffle convolution. Then, an attention mechanism composed of local attention and multi-scale attention fuses features from different meta-paths, resulting in the final patient embeddings. Finally, the model trains

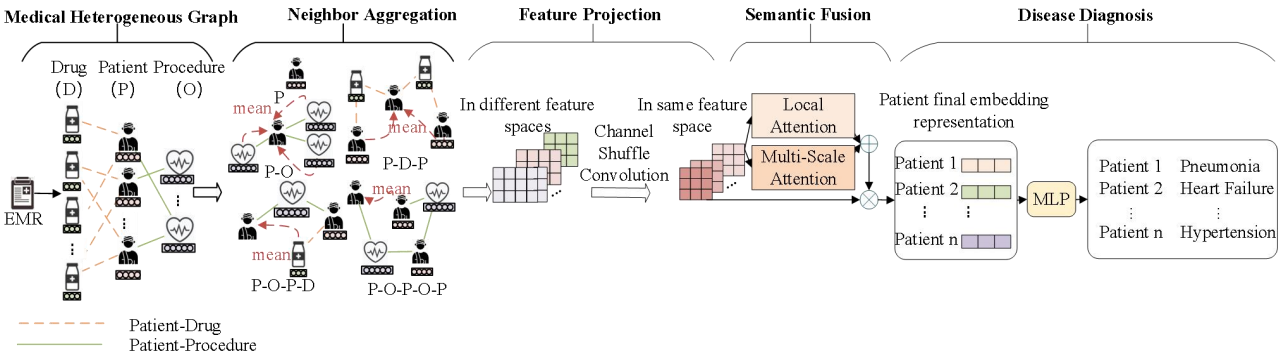

Fig.2. The overall architecture of MLM4DD

under supervised classification loss to predict the disease type of patients.

A. Medical Heterogeneous Graph Construction

The medical heterogeneous graph is constructed using the following approach. Using Fig.1 as an example, the constructed medical heterogeneous graph comprises three types of nodes: patients (P), drugs (D), and procedures (O). It includes two types of edges, namely patient-drug (indicating the drugs used by patients) and patient-procedure (indicating the procedures performed on patients). By connecting nodes with different edge relationships, the medical heterogeneous graph is constructed. These edge relationships reflect the associations between patients, drugs, and procedures. Specifically, given the medical heterogeneous graph $G=(V,E,A,R)$. The set of patients is denoted as $P=\{P_1,P_2,...P_k\}$, where k is the number of patients. The set of drugs is represented as $D=\{D_1,D_2,...,D_n\}$, where n is the number of drugs. The set of procedures is denoted as $O=\{O_1,O_2,...,O_m\}$, where m represents the number of procedures. Considering the adjacency matrix A of the medical heterogeneous graph. If a patient takes the medication D_n or undergoes the procedure O_m, the corresponding position in the adjacency matrix A is set to 1, otherwise, it is set to 0.

B. Neighbor Aggregation Based on Multi-Length Meta-Path Simplification

When applying attention and multi-layer structures of graph neural networks to disease diagnosis tasks, there are some limitations. Firstly, for large-scale medical heterogeneous graphs, introducing attention mechanisms may increase computational complexity, leading to decreased efficiency in information propagation and ultimately affecting diagnostic accuracy. Additionally, using multi-layer structures can make it challenging to distinguish high-level semantics, potentially causing issues like gradient vanishing or exploding, making model training difficult to converge and negatively impacting performance in disease diagnosis. To overcome these issues, inspired by the work in [4], a new approach is employed. This approach includes using an average aggregator to simplify neighbor aggregation and adopting a single-layer structure with a long meta-path. The simplified neighbor aggregation is performed only once during the preprocessing stage, generating a list $Z=\{X^M : M\in\Phi_X\}$ containing matrices with different semantic features for the set of given meta-paths. For each node V_i, an average aggregator is used to aggregate features based on the neighbor sets for each given meta-path, producing the list of the semantic feature vectors as follows:

$$Z_i=\{Z_i^M=\frac{1}{\|S^M\|}\sum\nolimits_{p(i,j)\in S^M}X_j : M\in\Phi_X\} \quad (1)$$

Here, S^M is the collection of all meta-paths instances corresponding to the meta-path M, while $M(i,j)$ is a meta-path instance containing the target node i and the source node j.

The path-based methods, such as HAN, enumerate neighbors based on each meta-path during preprocessing. This enumeration leads to an exponential increase in the number of meta-path instances as the length of the meta-path increases. Consequently, this results in high computational costs. Inspired by the hierarchical propagation of GCN, the simplification method reduces the set of neighbors based on meta-paths. It computes the final contribution weights for each node to the target using matrix multiplication with the adjacency matrix. This approach can handle large-scale medical heterogeneous graphs more efficiently, thus improving the model's training and inference processes. Specifically, assuming $X^c=\{x_0^{c^T};x_1^{c^T};...,x_{\|V^c\|-1}^{c^T}\}\in R^{\|V^c\|\times d^c}$ is the original feature matrix for all nodes, where $\|V^c\|$ is the number of nodes and d^c is the feature dimension. In this way, the simplified neighborhood aggregation process can be represented as follows:

$$X^M=\hat{A}_{c,c_1}\hat{A}_{c_1,c_2}...\hat{A}_{c_{l-1},c_l}X^{c_l} \quad (2)$$

Where $M=cc_1c_2...c_l$ is an l-hop meta-path, and $\hat{A}_{ci,ci+1}$ is the row-normalized form of the adjacency matrix $A_{ci,ci+1}$ between node types c_i and c_{i+1}. The aggregation result of short meta-paths can be seen as intermediate values for long meta-paths. By simplifying the neighborhood aggregation process, it becomes more efficient to compute features for short meta-paths, which can then be used as intermediate representations when constructing long meta-paths. For

example, given two meta-paths PDP and PPDP in the medical heterogeneous graph, you can first compute X^{PDP} and then compute $X^{PPDP} = \hat{A}_{PP} X^{PDP}$.

Furthermore, previous studies [21-22] have demonstrated that incorporating labels as additional input into the model can significantly enhance its performance. Therefore, label aggregation is introduced into the model. Similar to the aggregation of original features, using one-hot encoded labels allows them to propagate along various meta-paths, generating a series of matrices $\{Y^M: M \in \Phi_Y\}$. These matrices reflect the label distribution of corresponding meta-path neighbors, and the method of integrating label information enables the model to comprehensively consider both the features and labels of patients, thereby enhancing the representational capacity of patient nodes. As patient-type nodes are the target nodes, both ends of any meta-path $M \in \Phi_Y$ should be patient-type nodes M. Given a meta-path $M = c, c_1, c_2 ... c_{l-1} c \in \Phi_Y$, the label propagation process can be represented as follows:

$$Y^M = rm\_diag(\hat{A}^M)\ Y^c, \hat{A}^M = \hat{A}_{c,c1}\hat{A}_{c1,c2}...\hat{A}_{cl-1,c} \quad (3)$$

In the label matrix Y^c, for nodes in the training set, the corresponding rows adopt one-hot encoded label values, while other rows are filled with zeros. To prevent label leakage, diagonal values of the matrix multiplication result are removed, ensuring that each node does not receive its own true label information. Label propagation is performed in the neighborhood aggregation step, and the generated semantic matrix serves as an additional training input. The entire approach, through simplified neighbor aggregation and the use of a single-layer structure for long meta-paths, significantly enhances the efficiency and generalization ability of the model. Particularly when dealing with large-scale medical heterogeneous graph data, this method provides more accurate results for disease diagnosis tasks.

The integrated semantic matrix is as follows:

$$Z = \{X^M : M \in \Phi_X\} \cup \{Y^M : M \in \Phi_Y\} \quad (4)$$

This integration method further enhances the learning capabilities for patient nodes, enabling the model to adapt to different scales and complexities of medical heterogeneous graph data. Consequently, it achieves more accurate results in disease diagnosis tasks.

C. Feature Projection Based on Channel Shuffling Convolution

Due to the different dimensions or positions in different data spaces of semantic vectors from various meta-paths, it is necessary to project them into a common data space. The common practice is to define a semantically specific transformation matrix W^M for each meta-path M and compute the projected vector $H'^M = W^M X^M$. However, in this paper, channel shuffling is applied to the semantic vectors of different meta-paths first. Subsequently, a multi-layer convolution operation (conv) is employed to project data from different meta-paths into the same feature space. This enhances the modeling capability of the disease diagnosis model, allowing it to better capture complex features and associative information in patient data.

To better capture global complex feature representations, the channel shuffling technique is employed, introducing inter-channel information interaction. By applying channel shuffling to the semantic vectors of different meta-paths, the model is encouraged to learn richer cross-channel correlations, enhancing the model's expressive power for potential information in the medical heterogeneous graph. The calculation formulas are as follows:

$$Z'^M = Z^M_{B,C \bmod G,G,K} \quad (6)$$

$$Z''^M = Z'^M_{B,G,C \bmod G,K} \quad (7)$$

$$\hat{Z}^M = Z''^M_{B,C,K} \quad (8)$$

Where B is the batch size, C is the number of channels, G is the number of channel groups, and K is the embedding dimension.

Additionally, the introduction of multiple convolution layers (conv) includes normalization layers, non-linear layers, and dropout layers between two consecutive linear layers. The addition of these components helps the model capture and learn complex relationships in the heterogeneous medical graph more effectively, enhancing the model's ability to represent patient data. The calculation formulas are as follows:

$$H'^M = CONV_M(\hat{Z}^M) \quad (9)$$

D. Semantic Fusion Based on Local and Multi-Scale Attention

Leveraging attention mechanisms [23-25], the framework cleverly integrates semantic feature vectors from different meta-paths through a combination of local and multi-scale attention. This fine-tuned fusion captures key information in patient data more precisely, resulting in the generation of the final patient embedding. Semantic fusion based on local and multi-scale attention is illustrated in Fig.3.

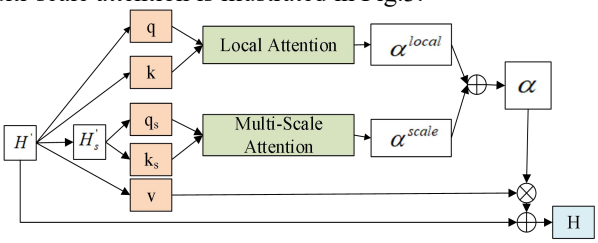

Fig.3. Semantic fusion based on attention

Firstly, by utilizing the local attention mechanism [24], the model focuses on specific neighborhood information along each meta-path. This enables the model to concentrate more on learning local features. It enhances sensitivity to each meta-path and ensures that details of each semantic are adequately addressed.

Specifically, using a predefined list of meta-paths $\Phi = \{M_1, M_2, ..., M_K\}$ and projected semantic vectors $\{h'^{M_1}, h'^{M_2}, ..., h'^{M_K}\}$, for each node, attention scores for each pair of semantic vectors are learned through local attention. Mapping each semantic vector h'^{M_i} to a query vector q^{M_i}, a key vector k^{M_i}, and a value vector v^{M_i}, the local attention scores $\alpha^{local}_{(M_i,M_j)}$ are computed as the dot product of the query vector q^{M_i} and the key vector k^{M_i}, normalized by the maximum and multiplied by the result of the window masking. The calculation formulas are as follows:

$$q^{M_i} = W_Q h^{'M_i},\ k^{M_i} = W_K h^{'M_i},\ v^{M_i} = W_V h^{'M_i},\ M_i \in \Phi \quad (10)$$

$$\alpha^{local}_{(M_i,M_j)} = \frac{\exp(q^{M_i} \cdot k^{M_j{}^T})}{\sum_{M_t \in \Phi} \exp(q^{M_i} \cdot k^{M_t{}^T})} \times window\_mask(M_j) \quad (11)$$

Where W_Q, W_K, W_V are trainable parameters shared across all meta-paths, and $Window\_mask(M_j)$ is the window mask that restricts the attention to a specific range.

Secondly, a multi-scale attention mechanism [25] directs attention towards information at different scales. This allows the model to learn across a broader semantic range. It enables the model to flexibly adjust the attention scope when dealing with various meta-paths, capturing global association information more effectively. Specifically, for each patient node, the model learns attention scores for each pair of semantic vectors using multi-scale attention. Depending on the scale, the input semantic vectors are scaled to obtain information at different scales $h_s^{'M_i}$, which is then mapped to query vectors q^{M_i} and key vectors k^{M_i}. When calculating multi-scale attention scores, attention weights $\alpha^{scale}_{(M_i,M_j)}$ for different scales are introduced and averaged, The calculation formulas are as follows:

$$h_s^{'M_i} = interpolate(h^{'M_i}, s) \quad (12)$$

$$q_s^{M_i} = W_Q^s h_s^{'M_i},\ k_s^{M_i} = W_K^s h_s^{'M_i},\ M_i \in \Phi \quad (13)$$

$$\alpha^{scale}_{(M_i,M_j)} = \frac{1}{S}\sum_{s=1}^{S} \frac{\exp(q_s^{M_i} \cdot k_s^{M_j{}^T})}{\sum_{M_t \in \Phi} \exp(q_s^{M_i} \cdot k_s^{M_t{}^T})} \quad (14)$$

$$\alpha_{(M_i,M_j)} = \alpha^{local}_{(M_i,M_j)} + \alpha^{scale}_{(M_i,M_j)} \quad (15)$$

$$h^{M_i} = \beta \sum_{M_j \in \Phi} \alpha_{(M_i,M_j)} v^{M_j} + h^{'M_i} \quad (16)$$

Where s represents the scale, $interploate()$ is the interpolation function, S represents the number of scales, and W_Q^s, W_K^s, β are trainable parameters shared across all meta-paths.

E. Disease Diagnosis

The final embedding for each patient node is the concatenation of all output vectors. The predictions for different disease labels are generated through a Multi-Layer Perceptron (MLP). The computation formula are as follows:

$$H = [h^{M_1} \| h^{M_2} \| ... \| h^{M_{|\Phi|}}] \quad (17)$$

$$\hat{\mathrm{y}} = MLP(H) \quad (18)$$

Where $\hat{y}_v \in R^C$ is the prediction, and C is the number of classes.

For a given set of training patients $\mathrm{V_{tr}}$, the overall loss is computed using cross-entropy with the following formula:

$$L = \sum_{v=V_{tr}} CROSSENT(\hat{y}_v, y_v) \quad (19)$$

Where $CROSSENT(\cdot)$ is the cross-entropy loss and $y_v \in R^C$ is the one-hot vector encoding the labels of the encoded node v.

V. Experiments and Evaluation

This section begins by introducing the datasets utilized in the experiments and the preprocessing steps applied to the data. Next, it delineates the evaluation metrics employed in the experiments and the baseline methods used for comparison. Subsequently, the performance of MLM4DD is detailed using experimental data.

A. Dataset and Preprocessing

The experiment utilized the MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset. This dataset covers clinical data from over 190,000 patients and 450,000 hospital admissions. In the data preprocessing stage, six representative disease categories were selected from the MIMIC-IV dataset. These categories include Myocardial Infarction, Pneumonia, Heart Failure, Coronary Atherosclerosis, Cirrhosis, and Hypertension. Key information was extracted from patient records. This included patient identifier (Subject_id), hospital admission identifier (Hadm_id), medication usage, medical procedures, gender, and disease diagnosis categories. Each patient has a unique Subject_id in the dataset, but they can correspond to multiple hospital admission records (multiple Hadm_id). To facilitate the processing of different patients, Subject_id and Hadm_id served as the primary keys for new patients. During patient selection, those with missing medication or procedure information were excluded. For medication usage, only drugs of the main (MAIN) type were selected while drugs of the base (BASE) type were removed. Additionally, a random selection of up to 30 drugs used by each patient was made. When processing patients' medical procedure data, only the most important procedures for each patient were chosen based on the importance ranking. Specifically, procedures with an importance ranking of 1, indicating the most vital procedure for the patient, were selected. After data preprocessing, the final dataset comprised 9,860 patients. The statistics of the processed dataset are summarized in Table II.

Table II
Statistics of Datasets

Disease label	Number of patients
Myocardial Infarction	1866
Pneumonia	1159
Heart Failure	2417
Coronary Atherosclerosis	2916
Cirrhosis	842
Hypertension	660
Total	9860

B. Evaluation Metrics

Micro-F1 and Macro-F1 are used as evaluation metrics for disease diagnosis tasks.

1) Micro-F1

Micro-F1 is an evaluation metric used in multi-class scenarios. It calculates the F1 score for each class and then computes their weighted average as an overall performance measure. The specific calculation is as follows:

$$Micro-F1=\frac{\sum_{i=1}^{n}2TP_i}{\sum_{i=1}^{n}(2TP_i+FP_i+FN_i)} \tag{20}$$

2) Macro-F1

Macro-F1 is an evaluation metric used in multi-class scenarios. It calculates the F1 score for each class and then computes their arithmetic average as an overall performance measure. The specific calculation is as follows:

$$Macro-F1=\frac{1}{n}\sum_{i=1}^{n}\frac{2TP_i}{(2TP_i+FP_i+FN_i)} \tag{21}$$

In which, n is the number of disease categories, and TP_i, FP_i, FN_i represent the counts of true positives, false positives, and false negatives for the i-th disease category, respectively.

C. Baselines

For a comprehensive evaluation of MLM4DD's performance, compare MLM4DD with the following baseline methods.

- GCN[26]: This method utilizes neighborhood aggregation operations to gather information from neighboring nodes in order to generate node representations.
- GAT [27]: This method employs an additional attention mechanism to achieve weighted aggregation of neighborhood information, instead of simple average aggregation.
- SlotGAT[12]: This method maintains representations in separate feature spaces for each node type by introducing independent slots for them. Slot attention is employed in the final layer to capture dependencies.
- Simple-HGN[14]: This method proposes a baseline model based on GAT, which simultaneously considers edge type embeddings and node embeddings to calculate attention scores.
- HINormer[28]: This method employs two key components enhanced by self-attention mechanisms to capture local and heterogeneous information in the graph, thereby facilitating node representation learning.
- HAN[16]: This method introduces a hierarchical attention mechanism. It is associated with meta-paths and implemented using GAT.

D. Parameter Setting

For GCN, GAT, SlotGAT, Simple-HGN, HINormer, and HAN, maintain the parameter settings according to their original papers and report the best performance.

The MLM4DD framework employs the Adam optimizer [29] throughout the training process. The learning rate is established at 0.001, the maximum hop count for long paths is set to 4, the channel shuffle grouping is defined as 3, the convolutional layers for feature projection are two, the scales for multi-scale attention are designated as 1 and 2, and there is a warm-up period of 100 epochs.

E. Experimental Results and Analysis

The experiment tests the performance of MLM4DD by completing the disease diagnosis task. MLM4DD uses 50% of the data as the training set, 20% as the validation set, and 30% as the test set. The findings drawn from the experimental results presented in Table III are as follows:

The proposed MLM4DD framework consistently demonstrates superior performance compared to other baseline methods. This indicates the effectiveness of combining local attention and multi-scale attention to learn representations of medical heterogeneous graphs, significantly improving the diagnostic performance of the framework. The performance of MLM4DD exceeds that of GCN, GAT, SlotGAT, Simple-HGN, and HINormer, demonstrating that the introduction of multi-length meta-paths can effectively capture the heterogeneity of medical heterogeneous graphs. The performance of MLM4DD surpasses that of HAN, indicating the effectiveness and necessity of simplified neighborhood aggregation and a single-layer structure with long meta-paths.

TABLE III
RESULTS OF DISEASE DIAGNOSIS EXPERIMENTS USING DIFFERENT METHODS

Model	Micro-F1	Macro-F1
GCN	82.01	82.36
GAT	85.59	85.86
SlotGAT	86.04	86.22
Simple-HGN	85.12	85.59
HINormer	86.82	87.07
HAN	81.54	80.86
MLM4DD	88.27	88.31

F. Variant Analysis

To validate the rationality of the MLM4DD framework structure, four variants of MLM4DD are proposed in this paper: MLM4DD_WOShuffle, MLM4DD_WOScale, MLM4DD_WOLocal, and MLM4DD_WOscale_local. MLM4DD_WOShuffle removes channel shuffling and uses regular convolution operations only. MLM4DD_WOScale removes multi-scale attention. MLM4DD_WOLocal removes local attention. MLM4DD_WOscale_local replaces multi-scale attention and local attention with the weighted and fused approach used in HAN. The performance on the MIMIC-IV dataset is compared with MLM4DD, and the experimental results are evaluated using Micro-F1 and Macro-F1, as shown in Fig. 4.

From this, the following conclusions can be drawn:

- MLM4DD_WOShuffle removes channel shuffling, which prevents cross-channel information interaction between meta-paths. Consequently, the framework is unable to learn richer cross-channel associations, leading to a decline in performance. This also indicates the necessity of channel shuffling for the MLM4DD framework.
- MLM4DD_WOScale removes multi-scale attention, which causes the framework to be unable to focus on relationships among nodes at different scales, leading to a decline in performance. This indicates that utilizing multi-scale attention allows the model to flexibly adjust the attention range when addressing various meta-paths, thereby better capturing overall associative information and enhancing the overall performance of the framework.
- MLM4DD_WOLocal eliminates local attention, resulting in the framework's inability to focus on specific

neighborhood information within each meta-path, which subsequently leads to a decline in performance. This signifies that local attention enables the model to better concentrate on learning local features, thereby effectively enhancing the model's sensitivity to each meta-path.

- MLM4DD_WOscale_local, which replaces multi-scale attention and local attention with HAN-weighted fusion, leads to a performance decline. This also indicates that the combination of local attention and multi-scale attention is superior to the HAN-weighted fusion approach.

In summary, the ablation experiments indicates the necessity of each component in the MLM4DD model.

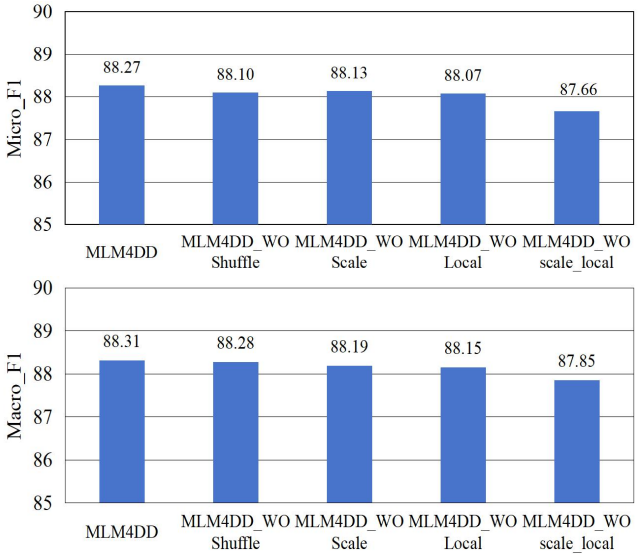

Fig. 4.The comparison of MLM4DD and its variant

G. Visualization

To evaluate the diagnostic results of the model, t-SNE [30] projects patient nodes from the test set into a two-dimensional space for visual analysis. The resulting visual representations are depicted in Fig. 5, where distinct colors denote labels corresponding to different disease categories.

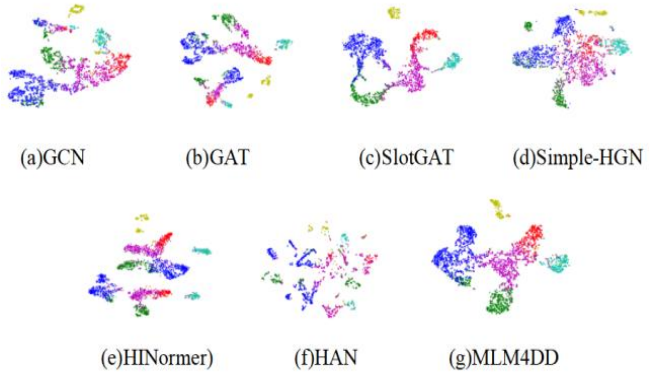

Fig.5. Visualization of the patient nodes embedding

Based on the experiments, it can be observed that in HAN, patient nodes with different labels do not aggregate well, while in GCN, patient nodes with different labels do not separate effectively. Other comparative learning methods, such as SlotGAT and HINormer, exhibit more distinct result boundaries and fewer overlapping regions but fail to aggregate nodes with the same label effectively. In the Simple-HGN method, nodes of different types are mixed together. In comparison to the aforementioned methods, the proposed MLM4DD method can effectively separate patient nodes with different labels and cluster patient nodes with the same label better. This indicates that MLM4DD can learn better patient node embeddings.

H. Parameter Analysis

This section discusses the parameter sensitivity analysis of MLM4DD, focusing on four important hyperparameters: the dimension of the hidden layer d, the number of convolutional layers for feature projection l, the dimension of local attention z, and the maximum length of meta-paths hop. By keeping other parameters constant and varying the values of d, l, z, and hop, we observe the performance changes of MLM4DD. Micor-F1 and Macro-F1 are used as evaluation metrics. Tables IV to VII respectively present the performance changes of MLM4DD under different dimensions of the hidden layer, numbers of convolutional layers for feature projection, dimensions of local attention, and maximum lengths of meta-paths.

a) Hidden layer dimension

As shown in Table IV, with the increase of the hidden layer dimension d, the performance of MLM4DD first decreases and then improves. The best performance is achieved when the embedding dimension is set to 512.

TABLE IV
THE PERFORMANCE VARIATION OF MLM4DD UNDER THE DIMENSIONS OF DIFFERENT HIDDEN LAYERS

Hidden layer dimension	Micro-F1	Macro-F1
64	86.85	86.73
128	87.15	87.18
256	87.42	87.41
512	88.27	88.31

b) The number of convolutional layers for feature projection

As shown in Table V, with the increase of the number of convolutional layers for feature projection, the performance of MLM4DD first increases and then decreases. The best performance is achieved when the number of convolutional layers for feature projection is set to 2.

TABLE V
THE PERFORMANCE VARIATION OF MLM4DD UNDER THE NUMBER OF CONVOLUTION LAYERS WITH DIFFERENT FEATURE PROJECTIONS

The number of convolutional layers for feature projection	Micro-F1	Macro-F1
1	87.96	88.08
2	88.27	88.31
3	87.96	88.02
4	87.73	87.79

c) Local attention dimension

As shown in Table VI, with the increase of the local attention dimension, the performance of MLM4DD first increases and then decreases. The best performance is achieved when the local attention dimension is set to 15 and 20. To improve efficiency, the local attention dimension is set to 15.

TABLE VI
THE PERFORMANCE VARIATION OF MLM4DD UNDER DIFFERENT LOCAL ATTENTION DIMENSIONS

Local attention dimension	Micro-F1	Macro-F1
5	88.24	88.27
10	88.20	88.26
15	88.27	88.31
20	88.27	88.31

d) Meta-path maximum length

As shown in Table VII, as the maximum length of meta-paths increases, the number of meta-paths also increases. The performance of MLM4DD first decreases and then increases. The best performance is achieved when the lengths of meta-paths cover 1, 2, 3, and 4. Therefore, the maximum length of meta-paths is set to 4.

TABLE VII
THE PERFORMANCE VARIATION OF MLM4DD UNDER DIFFERENT META-PATH MAXIMUM LENGTHS

Meta-path maximum length	Micro-F1	Macro-F1
1	88.03	87.92
2	88.00	87.95
3	87.90	87.93
4	88.27	88.31

I. Convergence Analysis

The loss function is one of the evaluation metrics during the training process of the MLM4DD model. The plot of the loss function versus the number of training iterations is shown in Fig. 6. It can be observed that at 100 epochs, the curve of the validation loss (val_loss) tends to stabilize and no longer decreases significantly, indicating that the model has converged.

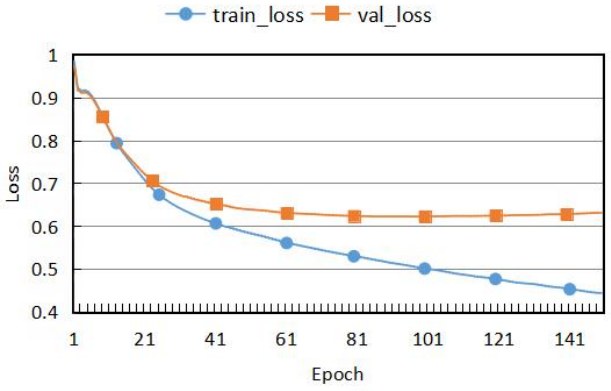

Fig.6 The Loss change curve

VI. CONCLUSIONS AND FUTURE WORK

In addressing the limitations of attention and multi-layer structures inherited from graph neural networks in the context of disease diagnosis tasks on heterogeneous graphs, a medical heterogeneous graph disease diagnosis framework, MLM4DD, based on multi-length meta-path semantic fusion is proposed. The framework utilizes electronic health records to construct a medical heterogeneous graph, employing a lightweight average aggregator for precomputing neighbor aggregation to reduce computational complexity and enhance information propagation efficiency. To better leverage semantic information and avoid issues such as vanishing or exploding gradients, MLM4DD introduces a single-layer structure with meta-paths of varying lengths to extend receptive fields. It incorporates local attention and multi-scale attention fusion to capture features from different meta-paths, thus obtaining embedded representations of patient nodes. Experimental results on the MIMIC-IV dataset demonstrate that MLM4DD outperforms baseline methods, effectively learning superior patient representations.

Future research directions will focus on how to leverage and integrate multimodal patient healthcare data, including medical textual information, X-rays, and other modalities. Integrating features from various modalities can lead to more accurate representations of patients, thereby further enhancing the performance of disease diagnostics.

REFERENCES

[1] Zhengkang Zhang, Dan Yang, and Yu Zhang, "Disease Diagnosis Based on Multi-View Contrastive Learning for Electronic Medical Records," IAENG International Journal of Applied Mathematics, vol. 53, no.3, pp1114-1122, 2023.

[2] Rushan Long, Dan Yang, and Yang Liu, "DiseaseNet: A Novel Disease Diagnosis Deep Framework via Fusing Medical Record Summarization," IAENG International Journal of Computer Science, vol. 49, no.3, pp808-817, 2022

[3] Zhihuang Lin, and Dan Yang, "Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity," Engineering Letters, vol. 28, no.3, pp651-662, 2020.

[4] Yang X, Yan M, Pan S, et al. "Simple and efficient heterogeneous graph neural network," in Proceedings of the AAAI Conference on Artificial Intelligence, vol.37, no.9, pp10816-10824, 2023.

[5] Hamilton W, Ying Z, Leskovec J. "Inductive Representation Learning on Large Graphs," Advances in Neural Information Processing Systems, 2017, 30. Available: https://arxiv.org/abs/1706.02216.

[6] Liu J, Song L, Wang G, et al. "Meta-HGT: Metapath-aware HyperGraph Transformer for heterogeneous information network embedding ," Neural Networks, vol. 157, pp65-76, 2023.

[7] Liu J, Song L, Gao L, et al. "MMAN: Metapath Based Multi-Level Graph Attention Networks for Heterogeneous Network Embedding (Student Abstract)," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no.11, pp13005-13006, 2022.

[8] Hong H, Guo H, Lin Y, et al. "An Attention-based Graph Neural Network for Heterogeneous Structural Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol.34, no.4, pp4132-4239, 2020.

[9] Yun S, Jeong M, Kim R, et al. "Graph transformer networks," in Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp11983-11993, 2019.

[10] Liu Z, Zheng V W, Zhao Z, et al. "Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding," in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp154-160, 2017.

[11] Zhu S, Zhou C, Pan S, et al. "Relation Structure-Aware Heterogeneous Graph Neural Network," in 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, pp1534-1539, 2019.

[12] Zhou Z, Shi J, Yang R, et al. "SlotGAT: slot-based message passing for heterogeneous graphs," in International Conference on Machine Learning, PMLR, pp42644-42657, 2023.

[13] Hong H, Guo H, Lin Y, et al. "An Attention-based Graph Neural Network for Heterogeneous Structural Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, vol.34, no.4, pp 4132-4139, 2020.

[14] Lv Q, Ding M, Liu Q, et al. "Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp1150-1160, 2021.

[15] Zhang C, Song D, Huang C, et al. "Heterogeneous graph neural network," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp793-803, 2019.

[16] Wang X, Ji H, Shi C, et al. “Heterogeneous Graph Attention Network,” in The World Wide Web Conference, pp2022-2032, 2019.
[17] Fu X, Zhang J, Meng Z, et al. “Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding,” in Proceedings of the Web Conference, pp2331-2341, 2020.
[18] Wang Z, Wen R, Chen X, et al. “Online Disease Self-diagnosis with Inductive Heterogeneous Graph Convolutional Networks,” in Proceedings of the Web Conference, pp3349-3358, 2021.
[19] Li Y, Zhao X, Ma M, et al. “VGBNet: a disease diagnosis model based on local and global information fusion,” International Journal of Computing Science and Mathematics, vol.17, no.2, pp107-122, 2023.
[20] Zheng S, Zhu Z, Liu Z, et al. “Multi-modal graph learning for disease prediction,” in IEEE Transactions on Medical Imaging, vol.41, no.9, pp2207-2216, 2022.
[21] Wang H, Leskovec J. “Unifying graph convolutional neural networks and label propagation,” in ArXiv Preprint ,2020. Available: https://arxiv.org/abs/2002.06755
[22] Wang Y, Jin J, Zhang W, et al. “Bag of tricks for node classification with graph neural networks,” in ArXiv Preprint, 2021. Available: https://arxiv.org/abs/2103.13355v3.
[23] Vaswani A, Shazeer N, Parmar N, et al. “Attention is all you need,” Advances in Neural Information Processing Systems, pp5998-6008, 2017.
[24] Fu P, Liu D, Yang H. “LAS-Transformer: An Enhanced Transformer Based on the Local Attention Mechanism for Speech Recognition,” Information, vol.13, no.5, pp2078-2489, 2022.
[25] Fu Y, Chen J, Zhang T, et al. “Residual scale attention network for arbitrary scale image super-resolution.” Neurocomputing, pp 201-211, 2021.
[26] Kipf T N, Welling M. “Semi-supervised classification with graph convolutional networks,” ArXiv Preprint, 2016. Available: https://arxiv.org/abs/1609.02907.
[27] Veličković P, Cucurull G, Casanova A, et al. “Graph attention networks,” ArXiv Preprint 2017. Available: https://arxiv.org/abs/1710.10903.
[28] Mao Q, Liu Z, Liu C, et al. “Hinormer: Representation learning on heterogeneous information networks with graph transformer,” in Proceedings of the ACM Web Conference 2023, pp 599-610, 2023.
[29] Kingma D P, Ba J. “Adam: A Method for Stochastic Optimization,” ArXiv: Learning,2014. Available: https://arxiv.org/abs/1412.6980v6.
[30] Laurens van der Maaten, Geoffrey Hinton. “Visualizing Data using t-SNE ,” Journal of Machine Learning Research. pp2579–2605, 2008.