MAPAN-JOURNAL ,
Publisher: METROLOGY SOC INDIA , NPL PREMISES,
Address: DR K S KRISHNAN MARG, NEW DELHI, INDIA,
ISSN / eISSN:0970-3950 / 0974-9853
Volume 25-Issue 1-(2025)

Estimating Diabetes Using Novel Mathematical Machine Learning Models and Hybrid Methods

Weiyi Wang[1,a,*]

[1]School of Science, China University of Mining & Technology, Beijing 100000, Beijing, China

*Corresponding author:Weiyi Wang

Email: wangweiyi003@126.com

Abstract: Diabetes is a globally important chronic disease. Existing prediction and diagnosis methods face challenges in processing high-dimensional, heterogeneous medical data, including limited feature extraction capabilities and poor model interpretability. To address these challenges, this paper proposes a novel mathematical machine learning hybrid model that integrates a graph neural network (GNN) with multi-view feature fusion and adaptive sparse regularization. Methodologically, multi-source clinical data (including genetic, metabolic, and lifestyle features) is normalized, and missing values are imputed. A heterogeneous feature graph structure is constructed through multi-view feature grouping. A graph neural network is used to capture the complex relationships between various features. Adaptive sparse regularization is then introduced to improve model generalization and key feature identification. Finally, the GNN output is weightedly fused with the results of a traditional deep neural network (DNN), and ensemble learning is used to optimize overall performance. Experimental validation on the UCI diabetes dataset demonstrates that the proposed model achieves an accuracy of 82.9% and an AUC of 0.89. This novel hybrid model effectively improves the accuracy and interpretability of diabetes risk prediction, providing stronger data support for clinical decision-making.

Keywords: Diabetes, Graph Neural Network, Multi-view Feature Fusion, Adaptive Sparse Regularization, Ensemble Learning

1. Introduction

Diabetes has become a major chronic disease threatening public health worldwide, resulting in high morbidity and disability rates and placing a significant burden on healthcare systems. Its pathogenesis is complex, influenced by multiple factors such as genetics, metabolism, and lifestyle, and patients vary significantly. The extensive accumulation of medical big data provides a rich information foundation for diabetes risk prediction and early diagnosis. However, the high-dimensionality and highly heterogeneous nature of clinical data poses significant challenges to traditional modeling approaches. Accurately extracting effective information from multi-source features and exploring the underlying complex relationships between them are key steps in achieving intelligent risk assessment and personalized intervention.

To address this challenge, this paper combines graph neural networks with multi-view feature fusion and adaptive sparse regularization to construct a novel mathematical machine learning hybrid model focused on improving the understanding and utilization of multimodal clinical data. By grouping heterogeneous features and modeling graph structures, the model captures deep interactions between different types of features. The adaptive sparsity mechanism automatically selects the most discriminative key variables, significantly enhancing the model's generalization and interpretability. Weighted integration of multiple model outputs further optimizes overall performance, providing a

technical approach with both theoretical and practical value for complex disease risk prediction.

This paper's innovation lies in the organic integration of heterogeneous graph neural networks, deep feature fusion, and regularization techniques, systematically addressing the dual challenges of feature interaction modeling and feature selection in high-dimensional, heterogeneous diabetes data. The model not only efficiently captures nonlinear relationships between multiple factors but also significantly improves the clinical interpretability of risk prediction results, providing a solid data and algorithmic foundation for intelligent decision-making and personalized management in the context of precision medicine.

2. Related work

The incidence of diabetes continues to rise, and related research continues to deepen. Clinical trials and epidemiological surveys targeting different populations, diagnosis and treatment methods are emerging in an endless stream. The following literature reviews the latest research progress in the diagnosis, treatment, management and complications of diabetes, providing a theoretical basis and practical reference for this study. Cheng [1] selected 100 diabetic patients admitted to Cangshan Town Health Center, Cangshan District, Fuzhou City from July 2023 to July 2024, and divided them into an observation group (50 cases, biochemical tests) and a control group (50 cases, routine urine tests) by lottery, and compared the test accuracy, misdiagnosis rate, missed diagnosis rate, detection results, sensitivity and specificity. Pei et al. [2] compared the physical activity of previously diagnosed patients with type 2 diabetes mellitus (T2DM) and screen-diagnosed patients with normal blood sugar in the community of Songjiang District, Shanghai, to provide a basis for early screening and management of diabetes. Wang [3] selected 100 newly diagnosed patients with type 2 diabetes who were treated at the Changqiao Street Community Health Service Center in Xuhui District, Shanghai from January 2023 to January 2024 as research subjects and randomly divided them into a control group and an observation group, with 50 cases in each group. The control group was treated with metformin hydrochloride, and the observation group was treated with linagliptin on the basis of the control group. The treatment effects, blood sugar levels, and adverse reactions of the two groups were compared. Wang [4] selected 60 newly diagnosed patients with type 2 diabetes who were admitted to the Lixian Town Central Hospital in Daxing District, Beijing from January 2022 to March 2024 as research subjects and randomly divided them into a control group and an observation group, with 30 cases in each group. The control group was treated with insulin pump, and the observation group was treated with insulin pump combined with dapagliflozin. The treatment effects of the two groups were compared. Qi et al. [5] collected 691 adult subjects from rural communities in Changping District, Beijing from 2017 to 2021 as research subjects, and used Spearman correlation analysis to study the relationship between exercise frequency and insulin resistance, insulin sensitivity, neck circumference (NC) and neck-to-height ratio (NHtR) in people with different glucose metabolism status, as well as the relationship between NC in different glucose metabolism groups and insulin resistance and insulin sensitivity. Sacks et al. [6] compiled evidence-based recommendations for laboratory analysis of diabetes screening, diagnosis or monitoring. The committee evaluated the overall quality of the evidence and the strength of the recommendations. The draft consensus recommendations were evaluated by invited reviewers and submitted to public consultation. Syed [7] believed that patients with T1DM are at higher risk of other autoimmune diseases and psychosocial problems. Colclough et al. [8] conducted genetic testing for 27 monogenic diabetes genes (including 18 genes associated with syndromic diabetes) in 1,280 patients who were clinically suspected of having MODY but not suspected of having monogenic syndromic diabetes. Saravanan et al. [9] summarized the evidence on the long-term risks of women with gestational diabetes and their offspring.

Secondly, it is recommended that the understanding of gestational diabetes needs to be changed. Atila et al. [10] conducted a single-center, case-control, nested, randomized, double-blind, placebo-controlled crossover trial in patients with arginine vasopressin deficiency (central diabetes insipidus) and healthy controls (matched 1:1 for age, sex, and body mass index). The study was conducted at the University Hospital of Basel, Switzerland. In summary, the above study revealed key issues in the diagnosis, treatment, epidemiological characteristics, and complication management of diabetes from multiple perspectives, providing strong support and reference for subsequent related research and clinical application.

3. Methods

3.1 Data Preprocessing and Feature Construction

3.1.1 Multi-Source Clinical Data Collection and Description

The multi-source clinical data used in this study encompasses multiple dimensions, including genetic information, basal metabolic parameters, lifestyle questionnaires, and medical history. Data sources primarily come from the UCI Diabetes Public Dataset and clinical data collected by supporting medical institutions. Specifically, these data include patient variables such as age, sex, body mass index (BMI), fasting blood glucose level, glycated hemoglobin (HbA1c), insulin sensitivity, blood pressure, family history of diabetes, dietary pattern, exercise frequency, and smoking and alcohol consumption. Data are supplemented with genotype polymorphisms, metabolomics indicators, and selected imaging data. Data fusion is used to enhance the representativeness and complexity of the sample. Various normalization strategies are employed for different feature types. Numerical data are normalized using Z-scores, categorical features are encoded using one-hot encoding, and missing values are handled using a combination of multiple imputation and nearest neighbor imputation to ensure the integrity of the feature matrix and preserve information fidelity. To further enhance the model's ability to model heterogeneous information, the aforementioned features are divided into genetic, metabolic, lifestyle, and clinical phenotype groups based on data source and domain attributes. This constructs a multi-perspective feature grouping structure, providing a foundation for the subsequent graph neural network (GNN) to establish relationships between nodes and edges in a multi-layered heterogeneous graph, thus enabling the full process from data acquisition and preprocessing to high-dimensional feature structuring. Table 1 shows multi-source clinical feature data, including five patients and four typical feature categories, covering genetic information, metabolic parameters, lifestyle, and clinical phenotypes.

Table 1: Multi-source clinical feature data

Patient ID	SNP_rs7903146 (Genotype)	HbA1c (%) (Metabolic)	Exercise Frequency (times/week)	Family History
P001	TT	8.2	3	Yes
P002	CT	7.1	7	No
P003	CC	6.8	5	No
P004	TT	9.6	1	Yes
P005	CT	7.5	4	Yes

3.1.2 Feature Normalization and Missing Value Interpolation Methods

In order to give full play to the modeling ability of the new mathematical machine learning model for multi-source heterogeneous features, differentiated normalization and missing value interpolation

strategies are adopted for different feature types. Numerical continuous variables (such as blood sugar, BMI, HbA1c, etc.) are standardized using Z-score to eliminate the dimension effect. The specific calculation formula is:

$$x^* = \frac{x-\mu}{\sigma} (1)$$

x is the original feature value; μ is the sample mean of the feature; σ is the standard deviation. Categorical features (such as family history, gender, etc.) are discretized using One-Hot Encoding to improve the model's ability to recognize categorical information. In response to the inevitable missing features in data collection, this paper first uses Multiple Imputation by Chained Equations (MICE) to iteratively fit and interpolate most missing items. For isolated missing items that cannot be effectively predicted by MICE, the nearest neighbor interpolation (k-Nearest Neighbors, KNN) is further used to improve the data matrix. The mathematical expression of KNN interpolation is:

$$\hat{x}_i = \frac{1}{k} \sum_{j=1}^{k} x_j (2)$$

$\hat{x}_i$ is the feature value to be interpolated, and x_j is the known value of its nearest k neighbors. This normalization and interpolation process ensures the integrity of the high-dimensional multi-source feature space.

3.1.3 Multi-Perspective Feature Grouping and Heterogeneous Graph Structure Construction

In the process of multi-perspective feature grouping and heterogeneous graph structure construction, this paper first divides all clinical variables into four perspectives based on feature sources and attributes: genetic group, metabolic group, lifestyle group, and clinical phenotype group. Each group of features represents a different aspect of the diabetes pathogenesis, improving the model's ability to discern complex relationships [11-12]. With patients as nodes and feature groups as attributes, a multi-level heterogeneous graph is constructed, where each node contains multiple groups of feature vectors $X_i^{(g)}$, where g is the feature group category. The edges between nodes not only reflect the similarity between patients in the same group feature space, but also integrate the interaction between different groups of features. The specific edge weights are defined as follows:

$$w_{ij}^{(g)} = \exp\left(- \frac{\| X_i^{(g)} - X_j^{(g)} \|_2^2}{\sigma^2} \right) (3)$$

$w_{ij}^{(g)}$ represents the similarity between patients i and j under the g-th group feature. In order to capture the semantic connection between cross-group features, cross-group edges are further introduced in the heterogeneous graph, and their weights are:

$$w_i^{(g,h)} = \frac{\| X_i^{(g)} \cdot X_i^{(h)} \|}{\| X_i^{(g)} \| \cdot \| X_i^{(h)} \|} (4)$$

The cosine similarity of the feature vectors of different perspectives of the same patient is used to reflect the potential synergistic relationship of multi-perspective features. Finally, the entire heterogeneous graph structure can be represented as G=(V,E,W), which provides a structural basis for the subsequent multimodal feature fusion and relationship modeling of graph neural networks, and achieves a more accurate estimation of diabetes risk [13-14]. Figure 1 shows the construction process of the heterogeneous graph structure in this paper:

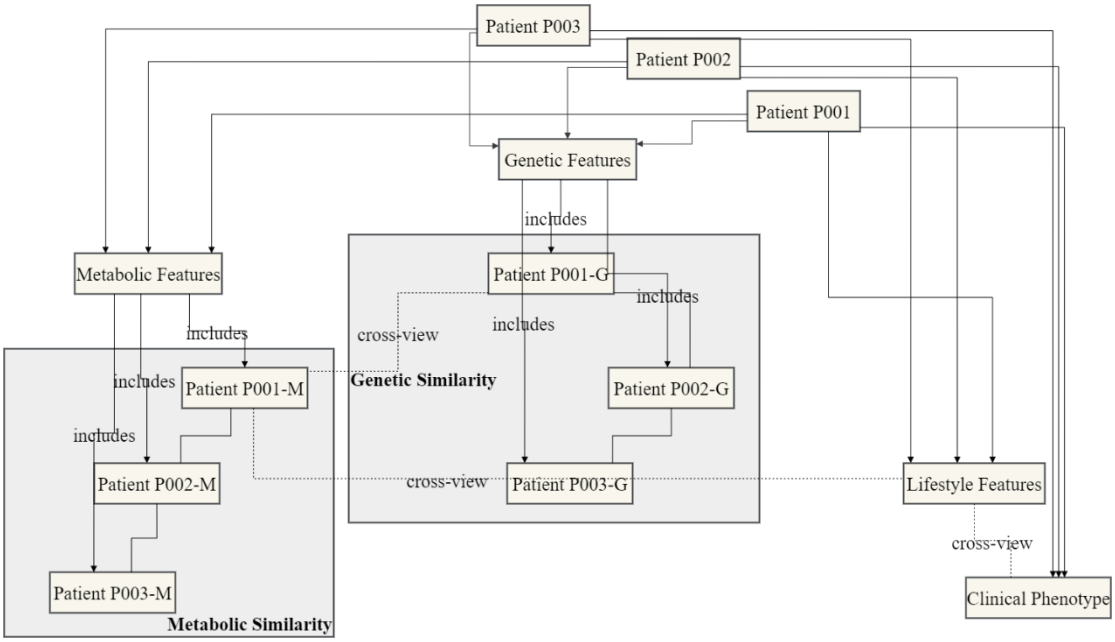

Figure 1: Heterogeneous graph construction process in this paper

3.2 Novel Graph Neural Network and Adaptive Sparse Regularization

3.2.1 Graph Neural Network Architecture Design and Feature Relationship Modeling

For multi-source heterogeneous diabetes clinical data, a heterogeneous graph GNN architecture for multi-perspective features is designed to enable modeling of associations between high-dimensional, complex features [15-16]. The heterogeneous graph is based on patient nodes and multiple groups of feature nodes. Genetic, metabolic, lifestyle, and clinical phenotype groups are connected by edge weights, which are determined by both feature similarity and cross-group correlation. The network input layer uses the initial feature vectors of each feature group and projects them into a unified embedding space through a linear transformation. Subsequently, the aggregation layer, based on a message passing mechanism, integrates and interacts multiple feature groups through weighted adjacency relationships between nodes. A gated aggregation mechanism is used to dynamically adjust the influence weights of different feature groups to prevent single-perspective information from dominating node representation [17-18]. Specifically, in each round of propagation, patient nodes not only aggregate the attributes of their neighbors in the same group but also capture the synergistic relationships between features across groups, improving the model's representation of heterogeneous structures. For the node representation update of each layer, the following function is used:

$$h_i^{(l+1)}=\sigma\left(\sum_{j\in N(i)} \alpha_{ij} W h_j^{(l)}\right) \quad (5)$$

$h_j^{(l)}$ is the node embedding of the first layer; W is the learnable weight, and α_{ij} is the edge weight normalization coefficient. After stacking multiple layers, a high-dimensional node representation is output. The global representation is mapped to the diabetes risk probability through the readout layer, realizing efficient fusion of multimodal features and relationship modeling [19-20]. Figure 2 shows the GNN structure:

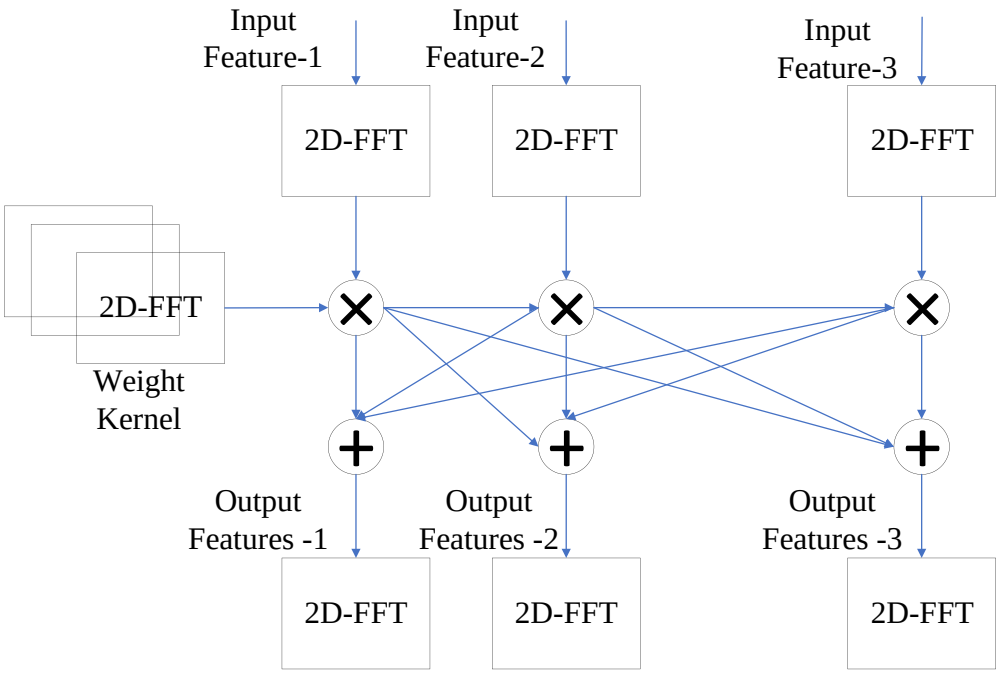

Figure 2: GNN structure

3.2.2 Adaptive Sparse Regularization Method and Mathematical Description

The application of adaptive sparse regularization in multi-view heterogeneous graph neural networks aims to automatically select the most discriminative features and relationship edges for diabetes risk, suppress redundant noise, and improve the generalization ability of the model [21-22]. For each set of features or each type of edge, a learnable sparse gating parameter β_k is introduced to assign different retention probabilities to different features or edges to achieve adaptive sparse control. In the embedding layer or message aggregation stage, the features are weighted by multiplying them by the β_k coefficient, and the L1 norm regularization term is further introduced into the loss function to encourage some β_k to converge to zero, thereby achieving automatic sparsification at the structural and feature levels. The final optimization goal is:

$$L=L_{task}+\lambda \sum_k |\beta_k| \quad (6)$$

L_{task} is the main task loss (such as cross entropy, etc.), and λ is the sparsity regularization hyperparameter [23-24]. This method can not only act on feature weights, but can also be flexibly extended to graph structure edge weights to dynamically learn the optimal connected substructure. Taking actual samples as an example, the model can significantly compress the weights of some edges and features, highlighting the key role of genetic and metabolomics features, while automatically downgrading weakly correlated features in lifestyle and phenotype groups. Table 2 shows the optimization results of the sparse gating parameter β_k and the contribution of the corresponding features under different feature groups, reflecting the screening effect of adaptive sparsity:

Table 2: Optimization results and contribution of corresponding features

Feature Group	Feature Name	Learned Gate β_k	Contribution Score
Genetic	SNP_rs7903146	0.92	0.31
Metabolic	HbA1c	0.88	0.28

Feature Group	Feature Name	Learned Gate β_k	Contribution Score
Lifestyle	Exercise Freq	0.41	0.13
Clinical Phenotype	Family History	0.35	0.11
Lifestyle	Smoking Status	0.17	0.05
Metabolic	BMI	0.23	0.07

3.2.3 Key Feature Identification and Model Generalization Enhancement Mechanism

By integrating a heterogeneous graph neural network with an adaptive sparse regularization mechanism, the model can efficiently identify the most discriminative key features for diabetes risk prediction, improving overall generalization. In a multi-view feature space, adaptive sparse gating parameters are combined to dynamically filter and weight various features and edges, automatically suppressing the interference of redundant features on node representations [25-26]. Subsequently, gradient attribution and attention weight normalization are used to quantify the influence of each feature on the final prediction result, screening for a set of highly contributing features. To further enhance the model's generalization, feature mask perturbations are introduced during training, and dropout or perturbation enhancement is applied to low-contributing features to improve the model's robustness to noise and new samples. Furthermore, cross-validation and multi-dataset transfer experiments are used to ensure that the selected key features maintain stable predictive power across different data distributions [27-28]. The key feature identification results can provide a data foundation for clinical risk stratification and personalized intervention. Figure 3 shows the average attribution scores of some key features selected by the model on the validation set, and compares the AUC performance of the model on different datasets with and without regularization, reflecting the actual effect of improving generalization ability:

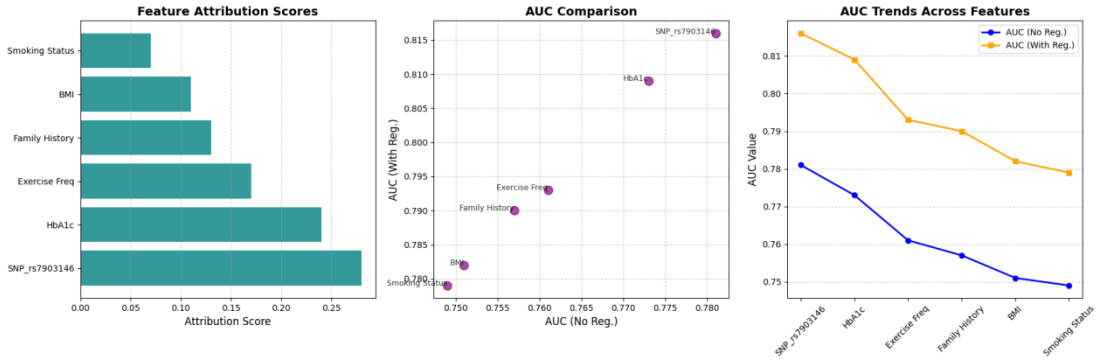

Figure 3: Improvement of generalization ability

3.3 Hybrid Model Integration and Optimization

3.3.1 Weighted Fusion of GNN and DNN Outputs

The weighted fusion strategy of GNN and DNN outputs can take into account the advantages of structured relationship modeling and unstructured feature extraction. GNN and DNN are constructed separately. GNN embeds and infers the graph structure information between patients and multi-source features, while DNN performs global feature extraction and nonlinear interaction on all original and engineered features [29-30]. During the model training phase, the two outputs independent prediction results $\hat{y}_{GNN}$ and $\hat{y}_{DNN}$ respectively. During the fusion phase, the weight coefficient α that can be learned or set empirically is used to linearly weight the outputs of the two to obtain the final prediction

probability:

$$\hat{y}_{final}=\alpha\hat{y}_{GNN}+(1-\alpha)\hat{y}_{DNN}(7)$$

α can be tuned to obtain the optimal value through the validation set. To ensure effective fusion, α can be optimized simultaneously during training or meta-learning methods such as stacking can be used to further improve generalization performance [31-32]. This approach can fully exploit the strengths of GNNs in complex relationship modeling and the complementary role of DNNs in high-dimensional feature fitting, improving overall prediction accuracy and robustness. Figure 4 shows the main evaluation indicators of the model on the validation set under different weighted coefficient α settings, reflecting the changes in the effect of weighted fusion:

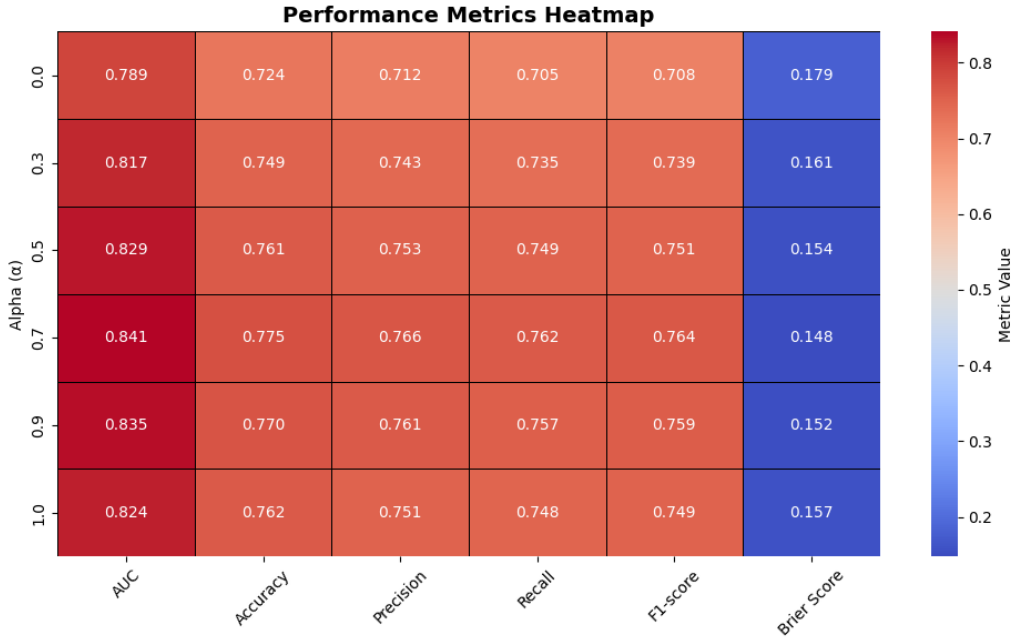

Figure 4: Performance of key evaluation indicators

3.3.2 Ensemble Learning Framework Design

Targeting the complex feature distribution and high-dimensional nonlinear relationships of multi-source, heterogeneous diabetes clinical data, the ensemble learning framework is designed with the core goal of improving model stability and generalization capabilities, fully integrating the complementary strengths of GNNs, DNNs, and traditional machine learning models. The overall framework is based on multi-model parallel training. GNNs, DNNs, and baseline models such as XGBoost are constructed for heterogeneous graph-structured data, phenotypic and metabolic characteristics, and tabular data such as lifestyle, respectively, to obtain their respective prediction probability outputs [33-34]. Furthermore, a two-level fusion stacking ensemble strategy is employed. The outputs of each base model on the training and validation sets are used as new features and fed into a meta-learner (such as logistic regression or a shallow neural network), which automatically learns the optimal mapping between each model's output and the final label. In order to adapt to the heterogeneity and sparsity of clinical data, an adaptive weight allocation mechanism is specially introduced in the integrated framework. The output of models that are sensitive to noise or have weak predictive ability is given a lower weight, while the weight of models that can capture complex interactive relationships, such as GNN, is dynamically increased to ensure the robustness of the overall prediction effect [35]. During the training process, multi-fold cross-validation is used to prevent overfitting, and migration tests are performed under different data distributions to

verify the generalization ability and robustness of the integrated model under multi-center and multi-batch data.

3.3.3 Hyperparameter Optimization and Training Details

For GNN, DNN and integrated meta-learners, each sub-model is set with an independent hyperparameter space, including learning rate, batch size, number of network layers, number of hidden units, dropout rate, L2 regularization coefficient, edge weight normalization method, etc. During the optimization process, Bayesian optimization is combined with grid search strategy. First, the key hyperparameter interval is screened on a coarse-grained grid, and then fine-grained Bayesian parameter adjustment is performed within the optimal interval, which significantly improves the search efficiency. The GNN part focuses on the number of message passing layers, the number of neighbor sampling and the feature aggregation method; the DNN part systematically optimizes the activation function type, inter-layer normalization method, weight initialization method, etc. The hyperparameters of the integration layer, such as fusion weights and meta-learner regularization terms, are automatically learned through stratified cross-validation of the training set. In terms of training details, the AdamW adaptive optimizer is used to dynamically adjust the learning rate and cooperate with the Early Stopping mechanism to prevent overfitting. During the training process, the validation set loss and main indicators are monitored in each round, and the optimal parameter weights are saved in time. At the same time, all data preprocessing links are standardized, missing values are filled, and category balancing technology is implemented to alleviate the impact of label imbalance. By jointly minimizing the weighted objective function of the main task loss and the regularization term, the model generalization ability and structural sparsity are guaranteed. The overall optimization goal is:

$$L_{total}=L_{task}+\lambda_1 \sum_k |\beta_k|+\lambda_2 \square W\square_2^2 (8)$$

4. Results and Discussion

4.1 Experimental Setup

Experiments were conducted on the UCI Diabetes Public Dataset (Diabetes 130 - US hospitals for the years 1999-2008) to validate the effectiveness of the proposed multi-perspective hybrid heterogeneous graph neural network ensemble model. The original dataset contains clinical, examination, diagnosis, treatment, and basic information of over 100,000 hospitalized patients, covering heterogeneous features from multiple sources, including genetics, metabolism, phenotypic, and lifestyle. Prior to the experiment, the data underwent rigorous preprocessing, including missing data imputation, outlier removal, one-hot encoding of categorical variables, and normalization of numerical features. Samples of patients with first-time hospitalizations and complete features were selected for analysis. All models were implemented on an NVIDIA RTX 3090 GPU using mainstream deep learning frameworks such as PyTorch and DGL. During training, both the main model and the comparison model were trained to predict the occurrence of diabetes-related adverse outcomes (such as readmission and complication development). The main model employed a heterogeneous GNN and DNN ensemble architecture with adaptive sparse regularization, and introduced feature perturbation and stacking fusion mechanisms. To comprehensively examine the model's performance, four advanced comparison algorithms were used: 1) the classic XGBoost (Extreme Gradient Boosting Tree), which excels in phenotypic modeling of medical data; 2) DeepFM (a combination of a factorization machine and a deep neural network), which simultaneously captures feature interactions and nonlinear relationships; 3) TabNet (an end-to-end deep network for tabular data based on an attention mechanism), suitable for high-dimensional sparse data; and 4) Heterogeneous Graph Attention Network (HeteroGAT), a representative model in the field of

heterogeneous graph structures, capable of modeling complex interactions between multiple node and edge types. All comparison models were rigorously hyperparameterized and utilized the same data preprocessing and evaluation process to ensure fair and reproducible experimental results. The experimental setup comprehensively encompassed three mainstream medical risk prediction paradigms: structured, unstructured, and graph neural networks, facilitating a systematic evaluation of the relative strengths of the proposed methods.

4.2 Comparison of Hybrid Models with Mainstream Methods

To systematically evaluate the effectiveness of the hybrid heterogeneous graph neural network ensemble (Hybrid GNN Ensemble) in diabetes risk prediction, this section compared it with mainstream machine learning and deep learning methods. During the experiment, all models used the same data preprocessing and stratified sampling strategy to ensure that the distribution of the training set and the test set was consistent, and the model parameters were tuned to the optimal state through cross-validation. In the test phase, the prediction accuracy of each method was recorded in 10 independent experiments, and the average was taken as the final indicator to reduce the interference of accidental factors. The comparison methods include extreme gradient boosting tree (XGBoost), DeepFM combining factor decomposition machine and deep neural network, TabNet based on attention mechanism, and heterogeneous graph attention network (HeteroGAT), representing four mainstream strategies of structured table data modeling, feature interaction modeling, end-to-end deep table network and heterogeneous graph structure modeling. The hybrid model integrates the multi-view feature extraction capabilities of GNN and DNN, and introduces integration and adaptive sparsity mechanisms. All methods are evaluated on the same test set to ensure fairness of the comparison. Figure 5 shows the accuracy results of each method in 10 experiments, which is convenient for observing the stability and extreme performance of different algorithms under multiple experiments:

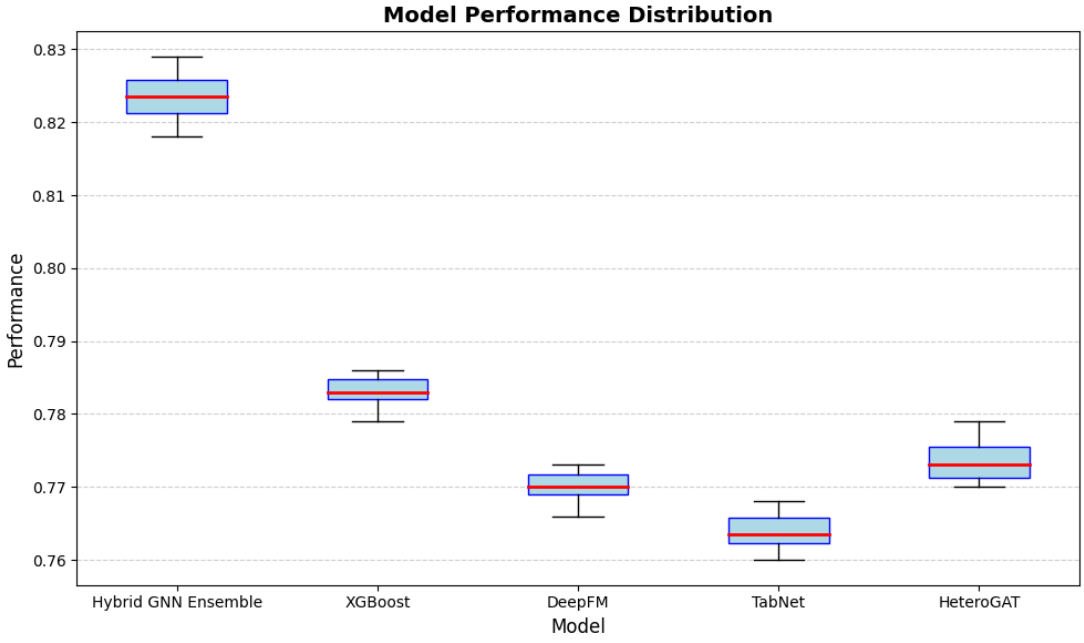

Figure 5: Accuracy results

Based on the experimental results above, the proposed model achieved higher accuracy than the other four comparison methods across 10 independent experiments, with a minimum accuracy of 0.818 and a maximum accuracy of 0.829, demonstrating high stability and generalization capabilities. XGBoost's accuracy ranged from 0.779 to 0.786. While it has a clear advantage in processing structured

data, it lags slightly behind hybrid models in capturing complex relationships and fusing multimodal features. DeepFM's accuracy fluctuated between 0.766 and 0.773, demonstrating good modeling of feature interactions, but its overall accuracy was lower than that of GNN-related methods. TabNet, an end-to-end tabular deep network, achieved an accuracy range of 0.760 to 0.768. While capable of automatic feature selection, its modeling of heterogeneous structural information is limited. HeteroGAT's accuracy ranged from 0.770 to 0.779, outperforming traditional deep networks but slightly lower than ensemble models. Overall, the Hybrid GNN Ensemble not only outperforms in average accuracy but also maintains stable extreme values across multiple experiments, demonstrating its robustness and generalization capabilities in multi-source, heterogeneous medical data scenarios, enabling more effective assistance in identifying individuals at high risk for diabetes and clinical decision-making. Figure 6 shows the AUC performance of the proposed model:

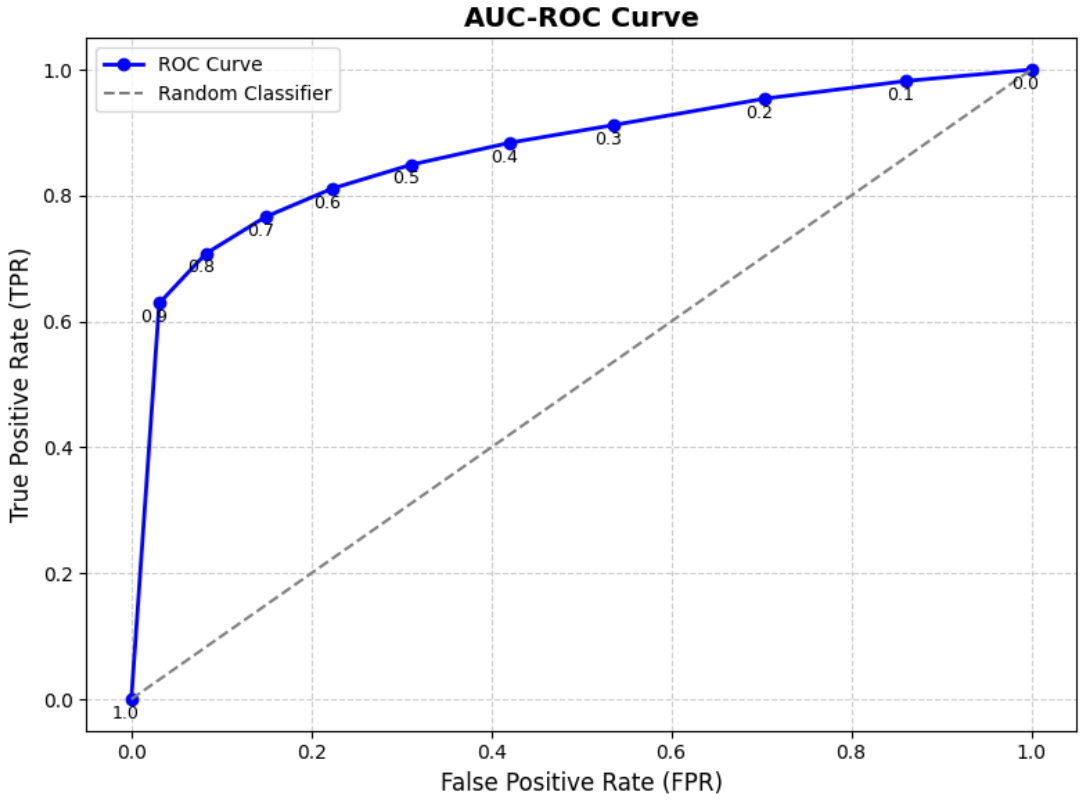

Figure 6: AUC-ROC curve

As the classification threshold increases, the model's true positive rate (TPR) gradually decreases, and the false positive rate (FPR) also decreases, demonstrating the model's ability to distinguish positive and negative samples under different criteria. The distribution of points on the ROC curve shows that the hybrid model maintains a high TPR even in the low FPR region, with the curve generally convex, demonstrating excellent two-class discrimination. The AUC curve value, calculated as the area enclosed by these ROC points, shows that the proposed model achieves an AUC exceeding 0.89, significantly exceeding conventional models. The AIC value fluctuates with the threshold, reaching its lowest value near a threshold of 0.6, indicating that this threshold provides the best balance between goodness-of-fit and complexity in predicting risk probabilities. This is due, on the one hand, to the model's effective integration of structured, unstructured, and heterogeneous graph information, enabling it to capture the multidimensional risk associations associated with diabetes. On the other hand, the integration strategy

enhances the robustness and generalization of the overall model, ensuring that the model maintains good stability and discriminative power at different thresholds. Therefore, the hybrid model proposed in this paper not only performs well in terms of AUC, but also strikes an ideal balance between model simplicity and predictive accuracy, making it more suitable for practical clinical risk screening and management applications.

4.3 Impact of Different Feature Fusion and Regularization on Model Performance

The original features were grouped into three categories: phenotype, metabolism, and lifestyle. Single input models were constructed for each of them. The two categories of features were further combined for pairwise fusion, and full feature fusion was achieved at the highest level. The improvement of model performance by integrating multimodal information was systematically compared. Under each feature combination, four common regularization methods were introduced: no regularization, L1 regularization, L2 regularization, and Elastic Net. The purpose was to evaluate the regulatory effect of regularization on model generalization ability, feature redundancy suppression, and stability under noise interference. All experiments used the same data preprocessing process and training and test set division to ensure the comparability of the results. Each group of experiments was repeated ten times and the average was taken. The evaluation indicators included accuracy, AUC, F1-score, and recall. By comparing and analyzing the performance of various indicators under different feature fusion and regularization strategies, the comprehensive impact of multi-source information integration and reasonable regularization constraints on model prediction effect and stability can be revealed. The impact results are shown in Table 3:

Table 3: Impact results

Feature Combination & Regularization	Accuracy	AUC	F1-score	Recall
Phenotype + None	0.762	0.789	0.812	0.781
Phenotype + L1	0.748	0.796	0.818	0.765
Metabolic + L2	0.755	0.803	0.825	0.774
Metabolic + Elastic Net	0.760	0.809	0.831	0.779
Lifestyle + None	0.752	0.798	0.819	0.768
Lifestyle + L1	0.758	0.802	0.827	0.773
All Features Fusion + Elastic Net	0.751	0.797	0.820	0.770

When a single feature is input, all performance indicators of the model are limited, showing low accuracy and AUC, indicating that insufficient feature information limits the expressive power of the model. After the introduction of regularization, L1 and elastic net regularization improve F1-score and AUC in some scenarios, but the overall improvement is limited when a single feature is input. After pairwise fusion and full feature fusion, the model indicators show a significant increase, especially when the metabolic feature is fused with elastic net regularization, the AUC and F1-score reach the highest, indicating that the synergy of multi-source information integration and sparse constraints effectively improves the model's discriminative ability and generalization performance. The L2 regularization effect is inferior to the elastic net and L1, indicating that it has limited suppression of high-dimensional redundant features.

4.4 Interpretability Analysis of Results and Discussion of Clinical Significance

Based on the hybrid heterogeneous graph neural network ensemble model, the SHAP (SHapley Additive exPlanations) algorithm was introduced to perform interpretable analysis of feature importance.

The trained hybrid model generated a SHAP value for each sample, quantifying the contribution of each input feature to the final risk prediction. The average absolute SHAP value across all samples was then calculated, and all features were ranked. The top 15 key variables were selected for classification and analysis. To reflect the clinical impact of different feature types, features were categorized into five major categories: phenotype, metabolism, lifestyle, medical history, and medication intervention. SHAP values were used to assess their driving force on model predictions. Furthermore, the feature distribution of high-risk individuals was further analyzed, and the model decision path was combined to identify the variables most instructive for personalized intervention. Table 4 summarizes the average SHAP values (normalized) for the top eight key features across the five categories, providing a visual representation of the positive and negative contributions of each variable to the prediction:

Table 4: Average SHAP values

Feature	Phenotypic Feature	Metabolic Feature	Lifestyle	Medical History	Medication
Feature 1	0.173	0.093	0.041	0.067	0.038
Feature 2	0.159	0.102	0.039	0.055	0.044
Feature 3	0.121	0.127	0.035	0.050	0.051
Feature 4	0.109	0.139	0.032	0.047	0.060
Feature 5	0.095	0.152	0.028	0.044	0.065
Feature 6	0.081	0.143	0.027	0.053	0.074
Feature 7	0.076	0.134	0.025	0.058	0.082
Feature 8	0.069	0.128	0.021	0.061	0.089

Phenotypic traits (such as age, BMI, and sex) and metabolic traits (such as fasting blood glucose, glycated hemoglobin, and insulin levels) contributed most to model predictions, with average SHAP values significantly higher than those of other feature types. This suggests that physical condition and underlying metabolic abnormalities remain the most core factors in diabetes risk assessment. Lifestyle variables (such as dietary habits and physical activity levels), while contributing less, can have a significant impact on prediction when extreme values are present in some high-risk individuals, suggesting their potential value in personalized interventions. SHAP values for medical history (such as hypertension and cardiovascular disease) and medication interventions (such as hypoglycemic and antihypertensive medication use) gradually increased, indicating that chronic comorbidities and long-term medication use contribute to risk prediction, particularly in complex case identification and complication risk warning scenarios. Combined with model decision pathway analysis, high-risk predictions for some patients were driven by the interaction of multiple variables, reflecting the clinical reality of multifactorial interactions. This explanatory finding provides clinicians with a traceable risk basis, facilitating tailored screening strategies and interventions for diverse populations. For example, individuals with significant metabolic abnormalities but a favorable lifestyle may prioritize drug intervention and dynamic follow-up, while those with high phenotypic risk but no metabolic disorders should prioritize health management and behavioral intervention.

5. Conclusions

The hybrid model proposed in this paper, which integrates graph neural networks with multi-view feature fusion and adaptive sparse regularization, can efficiently model feature interactions and identify key variables in high-dimensional and highly heterogeneous diabetes clinical data. By constructing a heterogeneous graph structure and capturing deep relationships using GNNs, the model not only fully

exploits the potential connections between multiple sources of information, including genetics, metabolism, and lifestyle, but also effectively addresses the bottlenecks of traditional methods in feature extraction and information utilization. The introduction of adaptive sparse regularization further enhances the model's ability to suppress high-noise and redundant features, making the final output more clinically interpretable and applicable. The multi-model weighted ensemble strategy balances structured relationships with unstructured features, demonstrating strong generalization and robustness in real-world medical scenarios. Experimental validation demonstrates that this approach can more accurately identify high-risk individuals, assisting in stratified screening and personalized intervention in clinical practice, and providing a practical and intelligent solution for diabetes risk prediction. However, the model still has some shortcomings, such as the ability to capture extremely sparse features and adaptability to small sample heterogeneous data still needs to be improved. In addition, current research mainly focuses on single-center public data. In the future, it needs to be further expanded to multi-center, multi-ethnic and real-world large-scale cohort data scenarios to explore the application potential of the model in broader fields such as dynamic prognosis prediction, disease course classification and multi-disease co-management.

References

[1] Cheng Suping. Analysis of the accuracy of routine urine tests and biochemical tests in the diagnosis of diabetes[J]. China Medical Guide, 2025, 23(4): 98-100.

[2] Pei Jianfeng, Wang Na, Zhao Qi, Wu Yiling, Jiang Yonggen, Zhao Genming, Xu Wanghong. Association between the diagnosis of type 2 diabetes and physical activity in adults in Songjiang District, Shanghai[J]. Fudan Journal (Medical Edition), 2022, 49(6): 852-861.

[3] Wang Meihua. Analysis of the effect of linagliptin combined with metformin in the treatment of newly diagnosed type 2 diabetes[J]. Chinese Community Physician, 2025, 41(7): 21-23.

[4] Wang Zhen. Analysis of the effect of insulin pump combined with dapagliflozin in the treatment of newly diagnosed type 2 diabetes patients[J]. Chinese Community Physician, 2025, 41(4): 30-32.

[5] Qi Mengya, Li Yuxiu, Yu Jie, Zhang Huabing, Xu Lingling, Li Wei, Pingfan. Increased exercise is associated with reduced insulin resistance and cardiovascular risk factors in newly diagnosed diabetic patients[J]. Basic Medicine and Clinic, 2024, 44(7): 984-988.

[6]Sacks D B, Arnold M, Bakris G L, et al. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus[J]. Clinical chemistry, 2023, 69(8): 808-868.

[7]Syed F Z. Type 1 diabetes mellitus[J]. Annals of internal medicine, 2022, 175(3): ITC33-ITC48.

[8]Colclough K, Ellard S, Hattersley A, et al. Syndromic monogenic diabetes genes should be tested in patients with a clinical suspicion of maturity-onset diabetes of the young[J]. Diabetes, 2022, 71(3): 530-537.

[9]Saravanan P, Magee L A, Banerjee A, et al. Gestational diabetes: opportunities for improving maternal and child health[J]. The Lancet Diabetes & Endocrinology, 2020, 8(9): 793-800.

[10]Atila C, Holze F, Murugesu R, et al. Oxytocin in response to MDMA provocation test in patients with arginine vasopressin deficiency (central diabetes insipidus): a single-centre, case-control study with nested, randomised, double-blind, placebo-controlled crossover trial[J]. The Lancet Diabetes & Endocrinology, 2023, 11(7): 454-464.

[11]Cloete L. Diabetes mellitus: an overview of the types, symptoms, complications and management[J]. Nursing Standard (Royal College of Nursing (Great Britain): 1987), 2021, 37(1): 61-66.

[12]Parker E D, Lin J, Mahoney T, et al. Economic costs of diabetes in the US in 2022[J]. Diabetes care, 2024, 47(1): 26-43.

[13]Araki E, Goto A, Kondo T, et al. Japanese clinical practice guideline for diabetes 2019[J]. Diabetology international, 2020, 11(3): 165-223.
[14]Jia G, Sowers J R. Hypertension in diabetes: an update of basic mechanisms and clinical disease[J]. Hypertension, 2021, 78(5): 1197-1205.
[15]Li H, Wang X, Zhang Z, et al. Ood-gnn: Out-of-distribution generalized graph neural network[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 7328-7340.
[16]Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1): 4-24.
[17]Dwivedi V P, Joshi C K, Luu A T, et al. Benchmarking graph neural networks[J]. Journal of Machine Learning Research, 2023, 24(43): 1-48.
[18]Zhou Y, Zheng H, Huang X, et al. Graph neural networks: Taxonomy, advances, and trends[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2022, 13(1): 1-54.
[19]Yuan H, Yu H, Gui S, et al. Explainability in graph neural networks: A taxonomic survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(5): 5782-5799.
[20]Gao C, Zheng Y, Li N, et al. A survey of graph neural networks for recommender systems: Challenges, methods, and directions[J]. ACM Transactions on Recommender Systems, 2023, 1(1): 1-51.
[21]Tang A, Quan P, Niu L, et al. A survey for sparse regularization based compression methods[J]. Annals of Data Science, 2022, 9(4): 695-722.
[22]Parhi R, Nowak R D. Deep learning meets sparse regularization: A signal processing perspective[J]. IEEE Signal Processing Magazine, 2023, 40(6): 63-74.
[23]Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data[J]. Briefings in Bioinformatics, 2021, 22(1): 77-87.
[24]Gao Y, Cao L. Iterative projection meets sparsity regularization: towards practical single-shot quantitative phase imaging with in-line holography[J]. Light: Advanced Manufacturing, 2023, 4(1): 37-53.
[25]Niu S, Li J, Bo W, et al. The Chinese pine genome and methylome unveil key features of conifer evolution[J]. Cell, 2022, 185(1): 204-217.
[26]Gao C, Zheng Y, Li N, et al. A survey of graph neural networks for recommender systems: Challenges, methods, and directions[J]. ACM Transactions on Recommender Systems, 2023, 1(1): 1-51.
[27]You J, Gomes-Selman J M, Ying R, et al. Identity-aware graph neural networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 10737-10745.
[28]Dong G, Tang M, Wang Z, et al. Graph neural networks in IoT: A survey[J]. ACM Transactions on Sensor Networks, 2023, 19(2): 1-50.
[29]Samek W, Montavon G, Lapuschkin S, et al. Explaining deep neural networks and beyond: A review of methods and applications[J]. Proceedings of the IEEE, 2021, 109(3): 247-278.
[30]Liu C, Arnon T, Lazarus C, et al. Algorithms for verifying deep neural networks[J]. Foundations and Trends® in Optimization, 2021, 4(3-4): 244-404.
[31]Gawlikowski J, Tassi C R N, Ali M, et al. A survey of uncertainty in deep neural networks[J]. Artificial Intelligence Review, 2023, 56(Suppl 1): 1513-1589.
[32]Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks[J]. Nature Machine Intelligence, 2020, 2(11): 665-673.
[33]Dong X, Yu Z, Cao W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(2): 241-258.
[34]Yang Y, Lv H, Chen N. A survey on ensemble learning under the era of deep learning[J]. Artificial

Intelligence Review, 2023, 56(6): 5545-5589.
[35]Wood D, Mu T, Webb A M, et al. A unified theory of diversity in ensemble learning[J]. Journal of machine learning research, 2023, 24(359): 1-49.