Hybrid GNN-Transformer Framework for Kidney Transplant Survival Prediction

Nazanin Rahmati
Department of Computer Science
Texas Tech University
Lubbock, TX 79409 USA
nrahmati@ttu.edu

Jingjing Yao, *Member, IEEE*
Department of Computer Science
Texas Tech University
Lubbock, TX 79409 USA
jingjing.yao@ttu.edu

***Abstract*—Accurate prediction of kidney transplant survival outcomes is essential for guiding personalized treatment decisions and improving long-term patient care. Traditional machine learning methods often fail to capture the complex, heterogeneous relationships and long-range dependencies within clinical data, limiting their predictive accuracy and generalizability. To address these challenges, this study introduces a hybrid Graph Neural Network–Transformer framework that integrates Graph Attention Network (GAT) and Graph Sample and Aggregate (GraphSAGE) layers to model fine-grained local and hierarchical patient relationships, followed by Transformer layers that capture global contextual dependencies across the entire cohort. Using real-world clinical data from 1,138 kidney transplant recipients obtained through the Hospital Information System of Iranian medical sciences university, the proposed hybrid model demonstrates consistently superior performance compared to existing approaches across multiple evaluation metrics, including F1 Score, Jaccard Index, and Precision–Recall Area Under the Curve (PRAUC). These results confirm that combining attention-based graph learning with Transformer-driven global context modeling significantly enhances prediction accuracy and generalization.**

***Index Terms*—Kidney transplant, graph attention network, graph neural network, transformer, classification.**

I. INTRODUCTION

Chronic kidney disease is a progressive loss of renal function that often leads to end-stage renal disease. It mainly affects individuals with diabetes, hypertension, or a family history of kidney dysfunction and is staged by the estimated glomerular filtration rate [1]. The rising global prevalence of kidney disease, driven by aging and metabolic disorders, poses a major healthcare burden. Kidney transplantation remains the most effective treatment, offering better survival, quality of life, and cost-efficiency than dialysis [2]. Yet, long-term graft survival remains uncertain due to rejection, infection, and non-adherence [3]. Accurate prediction of post-transplant outcomes is therefore vital for improving patient selection and organ allocation.

The availability of large, structured clinical datasets has enabled machine learning (ML) to uncover complex nonlinear relationships among demographic, biochemical, and immunological factors affecting graft outcomes. ML models now predict graft rejection, patient survival, and donor–recipient compatibility, offering insights beyond traditional statistical approaches [4]. However, most conventional ML methods treat clinical variables as independent features, ignoring relational dependencies among patients such as shared comorbidities or treatments. Because post-transplant success arises from interactions across multiple clinical dimensions, models that capture both individual and relational information are essential for reliable prediction.

Graph Neural Networks (GNNs) effectively model structured relationships in clinical data by representing patients as nodes and their similarities as edges [5]. This allows each node to learn not only from its own features but also from its neighbors, capturing both local and contextual information. Among GNN variants, Graph Attention Networks (GATs) assign adaptive weights to neighbors to emphasize clinically relevant relationships [6], while GraphSAGE [7] samples and aggregates neighborhood information to efficiently capture higher-order dependencies in large or evolving graphs [8]. Combining GAT and GraphSAGE thus balances fine-grained attention with scalable structural learning, making it well-suited for heterogeneous medical data.

While GNNs capture local structural dependencies, they struggle to model global relationships across distant or heterogeneous patients. To overcome this, we integrate Transformer layers, which use self-attention to capture cohort-wide dependencies beyond local neighborhoods [9]. In kidney transplant prediction, the Transformer enhances global context awareness, complementing the GNN's localized feature learning and enabling a unified representation of both micro- and macro-level patient relationships.

Based on this analysis, we propose a hybrid GNN–Transformer framework for kidney transplant survival prediction. After preprocessing with standard scaling, one-hot encoding, and Synthetic Minority Oversampling Technique (SMOTE) [10]. GAT and GraphSAGE layers learn multi-scale node embeddings, while a Transformer module captures global dependencies through self-attention across the cohort. This hybrid design integrates local, structural, and global contextual learning, resulting in higher predictive performance than existing models. The main contributions of this study are summarized as follows:

- We construct a comprehensive clinical dataset of kidney transplant recipients and design a three-stage preprocessing pipeline incorporating normalization, encoding, and oversampling to ensure balanced model training.
- We propose a hybrid GNN–Transformer model that integrates GAT, GraphSAGE, and Transformer layers to jointly learn local structural, neighborhood-level, and global contextual dependencies.
- We utilize a real-world dataset from the Hospital Information System of Iranian medical sciences university to evaluate our prediction model, demonstrating its improvement over existing approaches.

The rest of the paper is organized as follows. Section II reviews the related literature. Section III details the proposed framework, including data preprocessing, model architecture, and training methodology. Section IV presents simulation results and performance comparisons. Finally, Section V concludes the paper.

II. RELATED WORKS

Recent studies have increasingly applied ML to improve kidney transplantation survival prediction. Ravindhran *et al.* [11] conducted a comprehensive review of ML-based transplantation models and summarized commonly used architectures, including artificial neural networks (ANNs), convolutional neural networks (CNNs), decision trees (DTs), random forests (RFs), and hybrid ensemble methods for graft survival prediction. Paquette *et al.* [12] evaluated several survival models, including Cox proportional hazards regression, random survival forests, and deep neural networks such as recurrent neural networks (RNNs), for predicting graft survival probabilities. Naqvi *et al.* [13] applied ML-based classification models and deep autoencoders for dimensionality reduction to estimate graft failure risk across temporal cohorts. However, none of these works explicitly modeled relational structures or inter-patient dependencies within the clinical population.

Recurrent and attention–based models have been widely explored for clinical prediction tasks involving temporal dependencies. Wang *et al.* [14] proposed TAMSGC, which integrates temporal attention with simple graph convolution to model medication correlations over time. Choi *et al.* [15] developed RETAIN, a reverse-time attention model designed for interpretable visit-level analysis. Zhang *et al.* [16] introduced LEAP, which employs reinforcement learning for sequential decision-making to reduce adverse drug interactions. Yang *et al.* [17] designed SafeDrug with dual molecular graph encoders to capture drug structures and reduce unsafe combinations. However, none of the above recurrent or attention-based models considered graph-level patient similarity or long-range cohort dependencies, limiting their ability to model global clinical relationships.

GNN have been employed to capture structural dependencies in healthcare data. Shang *et al.* [18] developed GAMENet, which combines drug–drug interaction graphs with GCNs in a memory network to improve medication safety. Shang *et al.* [19] proposed G-BERT, which pre-trains graph-augmented Transformers to represent hierarchical medical code structures. Su *et al.* [20] introduced GATE, which constructs co-occurrence graphs with temporal updates to model visit-level relationships. Wang *et al.* [21] presented ARMR, leveraging adversarial regularization for sequential representation learning, and He *et al.* [22] proposed AMANet, a dual-view self-attention framework for asynchronous EHR sequences. Wang *et al.* [14] proposed PKANet, which incorporates heuristic diagnosis–medication mappings to guide attention-based learning. Gao *et al.* [23] introduced MK-GNN, which fuses drug network structures with EHR-derived patient features to improve medication representation and interpretability. However, these GNN-based approaches primarily focus on local neighborhood aggregation and lack Transformer-based mechanisms for modeling long-range cohort-level dependencies.

To the best of our knowledge, no prior study has applied GNN to kidney transplantation survival prediction. Existing GNN-based approaches in medical field have mainly focused on local neighborhood aggregation without modeling long-range dependencies across patients. The proposed hybrid GNN–Transformer framework overcomes these limitations by combining GAT and GraphSAGE for local and hierarchical feature learning, while Transformer layers capture global dependencies through self-attention. This unified design enables richer representations of clinical relationships and achieves higher predictive accuracy for kidney transplant survival.

III. FRAMEWORK DESIGN

The proposed hybrid GNN–Transformer framework, as shown in Fig. 1, is designed to predict kidney transplant survival outcomes by integrating heterogeneous clinical, demographic, and treatment data into a unified, graph-based learning architecture. The framework begins with structured input data preprocessing, where categorical and numerical patient attributes are normalized, one-hot encoded, and balanced using SMOTE to mitigate class imbalance. Each patient is represented as a node within a clinical graph, with edges encoding relational similarities such as shared medical histories, treatment patterns, or physiological characteristics. These node representations are then embedded into a latent feature space and processed through GAT and GraphSAGE layers, which jointly capture fine-grained neighbor importance and aggregated multi-hop structural information. The resulting node embeddings are passed through a Transformer-based global context module, which models long-range dependencies and inter-patient correlations across the entire cohort. Finally, the contextualized embeddings are fed into a fully connected classification layer that outputs the probability of graft survival.

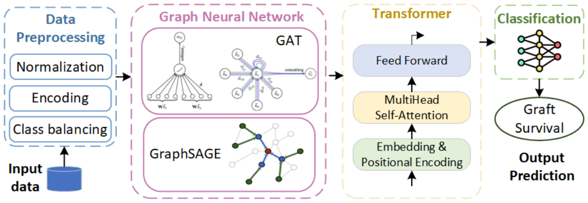


Fig. 1: Hybrid GNN–Transformer framework for kidney transplant outcome prediction.

A. Data Preprocessing

The kidney transplant dataset used in this study is obtained from the Hospital Information System of Iranian university of medical sciences. It contains 1,138 records of kidney transplant recipients, encompassing comprehensive demographic, socioeconomic, and clinical variables. Each record represents a unique patient profile with 25 features covering categories such as age, gender, blood group, primary cause of end-stage renal disease (ESRD), dialysis type, donor source, immunosuppressive drug usage, and post-transplant clinical indicators. The target label indicates transplant survival status, i.e., patients who survived after transplantation are labeled as 1 (successful), and those who did not are labeled as 0 (unsuccessful).

Before model training, the training dataset is processed through a structured pipeline to ensure consistency and suitability for graph-based learning. The preprocessing includes standardization of numerical attributes, one-hot encoding of categorical variables, and class balancing SMOTE.

1) Numerical Feature Standardization: Continuous attributes such as age, weight, ischemia time, and months from dialysis to transplant are standardized to remove scale disparities. Each numerical feature x_i is normalized using z-score normalization $x_i' = \frac{x_i - \mu_i}{\sigma_i}$, where μ_i and σ_i represent the mean and standard deviation of feature i, respectively. This ensures that all features have zero mean and unit variance, contributing proportionally during model training.

2) Categorical Feature Encoding: Categorical variables such as blood group, donor source, and insurance type are transformed into numerical form using one-hot encoding. For a categorical variable with C unique categories, each sample is represented as a binary vector $\mathbf{x}_i \in \{0, 1\}^C$, where a single element is set to 1 to indicate the active category. This transformation preserves categorical distinctions without introducing artificial ordinal relationships.

3) Class Balancing with SMOTE: The dataset contains 946 surviving and 236 non-surviving transplant recipients, resulting in class imbalance. SMOTE is applied to generate new minority samples through linear interpolation between each minority instance x_i and one of its k nearest neighbors $\tilde{x}_i$, i.e., $x_{\text{new}} = x_i + \lambda(\tilde{x}_i - x_i)$, where $\lambda \in [0, 1]$ is a random coefficient controlling the interpolation ratio. This approach expands the minority class and mitigates bias toward the majority class during model training.

B. Graph Neural Network

To model complex dependencies among kidney transplant patients, we construct a patient similarity graph and employ a hybrid GNN that integrates GAT and GraphSAGE layers. The GAT component adaptively learns the importance of neighboring patients, while GraphSAGE performs structural feature aggregation to ensure stability and generalization. This combination enables the model to capture both fine-grained local interactions and broader neighborhood-level patterns within clinical data.

1) Graph Construction: Each patient is represented as a node in an undirected graph $G = (V, E)$, where V denotes the set of N patients and E represents edges that connect clinically similar individuals. Edges are established using a k-nearest neighbor (k-NN) similarity measure applied to preprocessed feature vectors. The adjacency matrix A is defined as $A_{ij} = \begin{cases} 1, & \text{if } j \in \mathcal{N}(i) \\ 0, & \text{otherwise,} \end{cases}$ where $\mathcal{N}(i)$ denotes the set of node i's k-nearest neighbors. Each node i is initialized with a feature vector $h_i^{(0)} = x_i'$ obtained from the preprocessing stage.

2) Graph Attention Network (GAT) Layer: The GAT layer assigns adaptive importance to neighboring nodes, reflecting their clinical relevance. For node i and its neighbor $j \in \mathcal{N}(i)$, the unnormalized attention score is computed as $e_{ij}^{(l)} = \text{LeakyReLU}\Big(a^{(l)\top}\Big[W^{(l)}h_i^{(l)} \,\|\, W^{(l)}h_j^{(l)}\Big]\Big)$, where $W^{(l)}$ and $a^{(l)}$ are learnable parameters, and $\|$ denotes concatenation. The normalized attention coefficient is obtained using a softmax function $\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)}\exp(e_{ik}^{(l)})}$. The node representation is updated as $h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{(l)}W^{(l)}h_j^{(l)}\right)$, where $\sigma(\cdot)$ is a nonlinear activation. To stabilize training and capture diverse relational aspects, multi-head attention with M heads is used $h_i^{(l+1)} = \|_{m=1}^{M}\sigma\Big(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{(l,m)}W^{(l,m)}h_j^{(l)}\Big)$.

3) GraphSAGE Aggregation Layer: To complement the attention-driven focus of GAT and mitigate overfitting, the GraphSAGE layer aggregates neighborhood information using statistical averaging. Given the output $h_i^{(l)}$ from the previous GAT layer, the GraphSAGE update rule is defined as $h_i^{(l+1)} = \sigma\Big(W_1^{(l)}h_i^{(l)} + W_2^{(l)} \cdot \text{Avg}_{j\in\mathcal{N}(i)}h_j^{(l)}\Big)$, where $W_1^{(l)}$ and $W_2^{(l)}$ are learnable weights. This operation enforces smoothness and improves generalization across heterogeneous patient subgroups.

After the hybrid GAT–GraphSAGE block, the final node embeddings are expressed as $H^{(L)} = [h_1^{(L)}, h_2^{(L)}, \ldots, h_N^{(L)}]^\top$, where each $h_i^{(L)}$ encodes both local relational dependencies and smoothed statistical context. These embeddings are then passed to the Transformer layers for global context modeling.

C. Global Context Modeling with Transformer Layers

While the GNN module captures local and neighborhood-level dependencies, its receptive field is limited to a few graph hops. To model broader relationships across distant patient subgroups, such as variations in donor type, comorbidities, or immunosuppressive therapy, we incorporate Transformer layers that perform global self-attention over all patient embeddings.

Given the input embeddings $H^{(l)}$ at layer l, the queries Q, keys K, and values V are computed as $Q = H^{(l)}W_Q, K = H^{(l)}W_K, V = H^{(l)}W_V$, where W_Q, W_K, and W_V are learnable projection matrices, and d_k is the hidden dimension. The scaled dot-product attention mechanism is defined as $\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V$. Multi-head attention extends this process by concatenating M parallel attention heads. Each Transformer encoder layer applies residual connections, layer normalization, and a feed-forward network. Stacking L_T layers yields the globally contextualized embeddings $H^{(L_T)}$, where each patient representation aggregates information from all others, capturing cohort-level dependencies beyond the graph's local structure.

D. Final Classification and Model Training

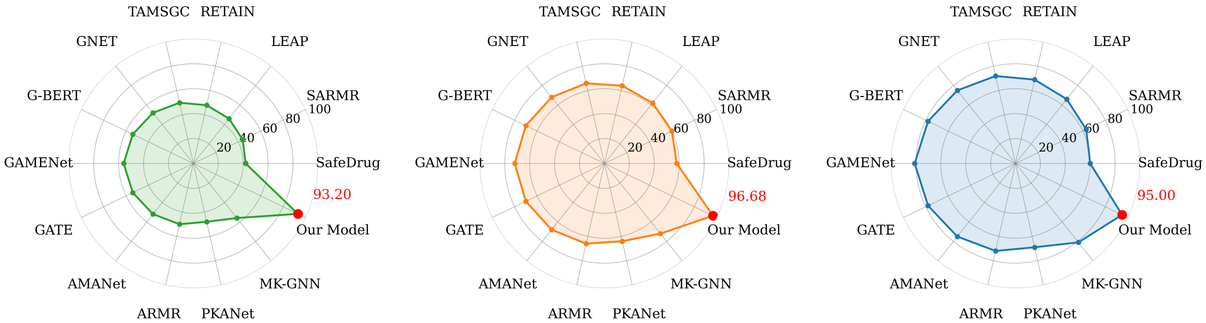
After GNN and Transformer processing, the model produces patient-specific embeddings that encode both local and global contextual information. Each embedding $h_i^{(L_T)}$ is passed through a fully connected prediction head followed by a Sigmoid activation to output the probability of transplant survival $\hat{y}_i = \sigma(W_{\text{final}}h_i^{(L_T)} + b_{\text{final}})$, where W_{final} and b_{final} are learnable parameters, and $\sigma(\cdot)$ maps the output to $[0, 1]$.

For binary classification, the model is trained using cross-entropy loss $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i\log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\Big]$, where $y_i \in \{0, 1\}$ is the ground-truth label. All parameters of the GNN, Transformer, and classification layers are jointly optimized via stochastic gradient descent to minimize $\mathcal{L}$, ensuring end-to-end learning of the complete framework.

IV. Performance Evaluation

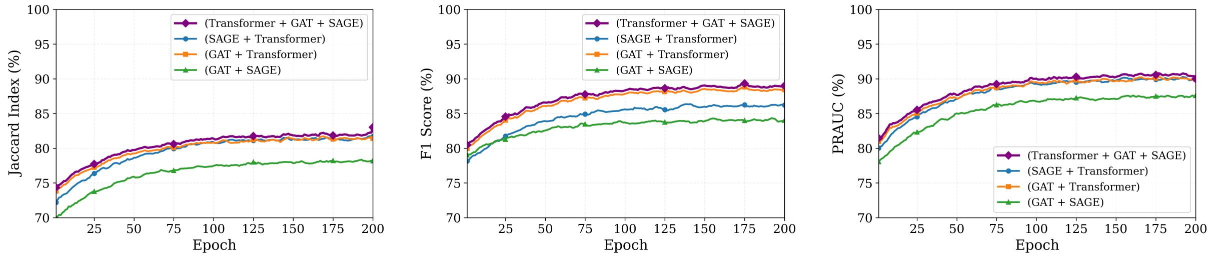
In this section, we evaluate the performance of the proposed kidney transplant survival prediction framework through simulations implemented in Python. We use a real-world dataset from the Hospital Information System of Iranian medical sciences university, containing medical records of 1,138 kidney transplant patients. Model performance is evaluated using three metrics including F1 Score, Jaccard Index, and Precision–Recall Area Under the Curve (PRAUC). The F1 Score measures the balance between precision and recall, the Jaccard Index quantifies the overlap between predicted and actual outcomes, and PRAUC assesses performance under class imbalance by capturing the trade-off between precision and recall. The proposed hybrid GNN-Transformer framework consists of two GAT layers with eight attention heads each, followed by two GraphSAGE layers, and then four Transformer encoder layers.

Fig. 2 compares the proposed hybrid GNN–Transformer model with representative baselines across Jaccard, F1 Score, and PRAUC metrics. RNN-based attention models, including TAMSGC [14], RETAIN [15], LEAP [16], SARMR [24], and SafeDrug [17], capture temporal dependencies but lack graph-based patient similarity modeling. Graph convolutional approaches, such as GAMENet [18], G-BERT [19], GATE [20], ARMR [21], AMANet [22], PKANet [14], and MK-GNN [23] incorporate patient graphs but omit Transformer-based global attention for long-range cohort reasoning. In Fig. 2(a), the hybrid GNN–Transformer achieves the highest Jaccard score by combining GAT and GraphSAGE



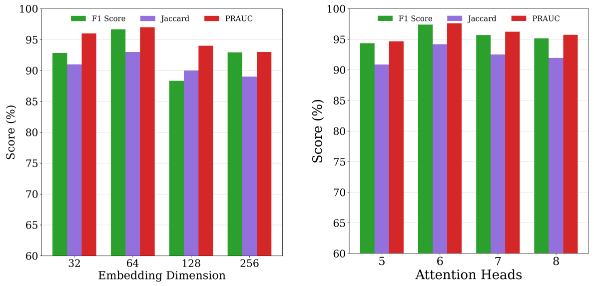
(a) Jaccard (b) F1 Score (c) PRAUC

Fig. 2: Performance comparison of our model with existing models.



(a) Jaccard (b) F1 Score (c) PRAUC

Fig. 3: Performance comparison of architectural variants.



(a) Embedding dimension (b) Attention heads

Fig. 4: Performance comparison with different hyper-parameters.

to capture both fine-grained edge importance and broader structural relationships, improving alignment between predictions and ground truth. Fig. 2(b) shows superior F1 performance, as the integrated Transformer enhances precision and recall by modeling global contextual dependencies missed by RNN-based and local GNN models. Fig. 2(c) demonstrates the highest PRAUC, indicating strong discrimination under class imbalance, as the multi-scale attention design enables the model to focus on clinically significant minority cases.

Fig. 3 illustrates the epoch-wise performance of four architectural variants (Transformer + GAT + GraphSAGE, SAGE + Transformer, GAT + Transformer, and GAT + SAGE), evaluated across Jaccard, F1 Score, and PRAUC metrics over 200 training epochs. The complete hybrid model (Transformer + GAT + GraphSAGE) consistently achieves the highest performance and fastest convergence across all metrics. In Fig. 3(a), the Jaccard results demonstrate that integrating the Transformer with both GAT and GraphSAGE improves prediction consistency by combining attention-based edge weighting with multi-hop neighborhood aggregation. Fig. 3(b) shows that the hybrid model attains superior F1 scores, as Transformer layers enhance precision–recall balance by capturing global dependencies beyond local graph propagation. In Fig. 3(c), the PRAUC trend highlights the model's strong discriminative ability and training stability under class imbalance.

Fig. 4 presents the parameter sensitivity analysis of the hybrid GNN–Transformer model with respect to two key hyperparameters: embedding dimension and number of attention heads. Fig. 4(a) examines how varying the embedding dimension from 32 to 256 influences model performance across F1 Score, Jaccard, and PRAUC. The results show that performance improves as the embedding dimension increases up to 64, reflecting richer and more discriminative feature representations, but declines beyond this point due to overfitting and redundant information. Fig. 4(b) illustrates the effect of changing the number of attention heads on model accuracy. Moderate head counts of 6 yield the best results for F1 Score and Jaccard, demonstrating that multi-head attention effectively captures diverse relational patterns among patients, whereas excessive heads fragment the embedding space and reduce stability. PRAUC remains relatively stable across head variations, suggesting the model's robustness under class imbalance. Overall, these results identify the optimal configuration for kidney transplant survival prediction.

V. CONCLUSION

In this paper, we have presented a hybrid GNN–Transformer framework for predicting kidney transplant survival outcomes using heterogeneous clinical, demographic, and treatment data. We have integrated GAT and GraphSAGE layers to jointly capture fine-grained patient-level relationships and hierarchical neighborhood structures, while Transformer layers model long-range dependencies and cohort-level interactions through self-attention. This unified design enables the model to learn comprehensive patient representations that capture both local and global contextual information. Simulation results have demonstrated that the proposed hybrid architecture consistently outperforms existing models across metrics F1 Score, Jaccard Index, and PRAUC. The analysis of architectural variants and parameter sensitivity further confirms that the combination of GAT, GraphSAGE, and Transformer components is essential for achieving optimal performance.

VI. ACKNOWLEDGEMENT

All the procedures of the study follow the Helsinki declaration for human procedures and experiments. Before the enrollment of the subjects, the aim of the study and the procedure was described for patients' parents, and signed informed consent was obtained for each patient for study participation. The feasibility of the study has been endorsed by the ethics committee of Shahid Beheshti University of medical sciences (Reference IR.SBMU.RETECH.REC.1403.842).

REFERENCES

[1] P. Romagnani *et al.*, "Chronic kidney disease," *Nat. Rev. Dis. Primers*, vol. 11, no. 1, p. 8, 2025.

[2] K. Panda and A. Mazumder, "Predicting dosage of immunosuppressant drugs after kidney transplantation using machine learning," in *Proc. IEEE URTC*. IEEE, 2023, pp. 1–5.

[3] Y. Hong *et al.*, "Improved waitlist and comparable post-transplant outcomes in simultaneous heart-kidney transplantation under the 2018 heart allocation system," *J. Thorac. Cardiovasc. Surg.*, vol. 167, no. 3, pp. 1064–1076, 2024.

[4] I. Furtney, R. Bradley, and M. R. Kabuka, "Patient graph deep learning to predict breast cancer molecular subtype," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 5, pp. 3117–3127, 2023.

[5] V. Ponzi and C. Napoli, "Graph neural networks: Architectures, applications, and future directions," *IEEE Access*, 2025.

[6] L. Sun *et al.*, "Class-driven graph attention network for multi-label time series classification in mobile health digital twins," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3267–3278, 2023.

[7] T. Liu *et al.*, "Graphsage-based dynamic spatial–temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 11 210–11 224, 2023.

[8] S. Lee, Y.-S. Moon, and H.-J. Jang, "Domain transformation to graphs and graphsage-based embedding for performance enhancement in time-series classification," *IEEE Access*, 2024.

[9] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[10] D. Dablain, B. Krawczyk, and N. V. Chawla, "Deepsmote: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6390–6404, 2022.

[11] B. Ravindhran *et al.*, "Machine learning models in predicting graft survival in kidney transplantation: meta-analysis," *BJS open*, vol. 7, no. 2, p. zrad011, 2023.

[12] F.-X. Paquette *et al.*, "Machine learning support for decision-making in kidney transplantation: Step-by-step development of a technological solution," *JMIR Med. Inform.*, vol. 10, no. 6, p. e34554, Jun 2022.

[13] S. A. A. Naqvi *et al.*, "Predicting kidney graft survival using machine learning methods: Prediction model development and feature significance analysis study," *J. Med. Internet Res.*, vol. 23, no. 8, p. e26843, Aug 2021.

[14] H. Wang *et al.*, "Medication combination prediction using temporal attention mechanism and simple graph convolution," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 10, pp. 3995–4004, 2021.

[15] E. Choi *et al.*, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[16] Y. Zhang *et al.*, "LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity," in *Proc. ACM SIGKDD*, 2017, pp. 1315–1324.

[17] C. Yang *et al.*, "Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations," *arXiv preprint arXiv:2105.02711*, 2021.

[18] J. Shang *et al.*, "Gamenet: Graph augmented memory networks for recommending medication combination," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 1126–1133.

[19] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," *arXiv preprint arXiv:1906.00346*, 2019.

[20] C. Su, S. Gao, and S. Li, "Gate: graph-attention augmented temporal neural network for medication recommendation," *IEEE Access*, vol. 8, pp. 125 447–125 458, 2020.

[21] Y. Wang *et al.*, "Adversarially regularized medication recommendation model with multi-hop memory network," *Knowl. Inf. Syst.*, vol. 63, pp. 125–142, 2021.

[22] Y. He *et al.*, "Attention and memory-augmented networks for dual-view sequential learning," in *Proc. ACM SIGKDD*, 2020, pp. 125–134.

[23] C. Gao *et al.*, "Medical-knowledge-based graph neural network for medication combination prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.

[24] Y. Wang *et al.*, "Self-supervised adversarial distribution regularization for medication recommendation." in *Proc. IJCAI*, 2021, pp. 3134–3140.