A Patient Similarity Graph Neural Network for Cardiovascular Disease Prediction

Dahiru Sulaiman Rashid*, Mohammad Hafiz Danmanu†, Mudathir Muhammad Salahudeen‡, Sulaiman Muhammad Rashid§

*Department of Mathematics, Gombe State University, Gombe, Nigeria
†Monitoring and Evaluation Research and Learning Unit, Achieving Health Nigeria Initiative, Adamawa, Nigeria
‡Data Analytics Department, Natview Foundation for Technology Innovation, FCT, Nigeria
§Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju, S/Korea
Email: dahirslmn@gmail.com

***Abstract*—Cardiovascular disease (CVD) remains the leading cause of death globally. Early detection from routine clinical records is challenging due to heterogeneous tabular features and class imbalance. This work presents a Patient Similarity Graph Neural Network (PS-GNN) that frames CVD prediction as a node classification problem on a patient similarity graph. The proposed pipeline first projects tabular patient features via a Multilayer Perceptron (MLP) to obtain latent embeddings, then applies stacked Graph Attention Network (GAT) layers to aggregate cohort-level information. To mitigate dataset skew, we combine Synthetic Minority Over-sampling Technique (SMOTE) with a class-weighted cross-entropy loss. We evaluate PS-GNN on three benchmark datasets (UCI Heart Disease, Cleveland Clinic, Framingham Heart Study) and report consistent improvements in macro F1 and Area Under the ROC Curve (AUC): UCI — Accuracy 0.82, Cleveland — Accuracy 0.85, Framingham — Accuracy 0.85. The results indicate that combining local feature encoding and graph-based relational learning yields robust and clinically useful predictions. All abbreviations (PS-GNN, MLP, GAT, SMOTE) are expanded on first use.**

***Index Terms*—Cardiovascular Disease, Graph Neural Network, Patient Similarity, Multilayer Perceptron, Graph Attention Network, SMOTE.**

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, responsible for millions of deaths annually [1]. Early detection and risk stratification are pivotal for effective clinical intervention. While traditional machine learning (ML) techniques such as decision trees, support vector machines, and linear discriminant analysis have been applied for CVD risk prediction, they commonly process each patient independently and do not exploit inter-patient relationships [3]–[5]. This independence assumption limits their ability to model cohort structure and latent correlations present in clinical populations.

Graph Neural Networks (GNNs) provide a flexible mechanism to model such relationships by representing patients as nodes and similarity relations as edges in a graph. Recent studies have demonstrated the potential of GNNs in healthcare settings where relational structure improves predictive performance [6], [7]. Motivated by these findings, we design a compact hybrid architecture that projects tabular features to a latent space using an MLP and applies attention-based graph aggregation (GAT) to leverage cohort information — the resulting model is the Patient Similarity Graph Neural Network (PS-GNN).

- We present PS-GNN, a hybrid MLP + GAT architecture tailored for tabular clinical datasets, which frames CVD prediction as node classification on a patient similarity graph.
- We propose a practical patient similarity graph construction procedure and robust training pipeline combining SMOTE with class-weighted loss.
- We validate the approach across three benchmark CVD datasets, provide a detailed dataset summary, and compare PS-GNN with conventional and graph-based baselines.

2. Related Works

Recent advances in graph-based learning for biomedical applications show promising improvements over flat models. Ochoa and Mustafa [6] review graph representations for clinical data and show how relational modeling helps in prognosis and treatment prediction. Paul et al. [7] provide a systematic review focusing on the use of GNNs in healthcare and emphasize attention mechanisms for interpretability. Prior works on CVD prediction have explored classical ML techniques [3]–[5] and deep-learning variants; however, many studies are limited to single-dataset evaluations or lack cross-cohort validation. Our work builds on these directions by combining MLP-based feature extraction with attention-based graph message passing and evaluating the method across multiple public datasets.

3. Methodology

This section describes data preparation, patient similarity graph construction, model architecture, and training.

3.1. Problem formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote N patient records where $\mathbf{x}_i \in \mathbb{R}^d$ is the tabular feature vector and y_i is the class label. The task is to learn $f : \mathbb{R}^d \to \mathcal{Y}$ that predicts y_i while exploiting inter-patient relations encoded in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

3.2. Data preprocessing and balancing

Categorical variables are one-hot encoded and continuous features standardized via z-score normalization. To address class imbalance, we apply the Synthetic Minority Oversampling Technique (SMOTE) on the training set to augment minority class examples; in addition, a class-weighted loss is used during optimization.

3.3. Patient similarity graph construction

We form an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $v_i \in \mathcal{V}$ denotes a patient. Edges are defined using a k-nearest neighbors (k-NN) rule in the normalized feature space. Formally, an edge is present if:

$$(v_i, v_j) \in \mathcal{E} \iff \mathbf{x}_j \in \mathrm{kNN}(\mathbf{x}_i). \tag{1}$$

Equation (1) enables contextual information to propagate among clinically similar patients.

3.4. Proposed architecture (PS-GNN)

Figure 1 illustrates the PS-GNN architecture (MLP projection → GAT layers → classification head). We first introduce the projection stage and its equation.

MLP projection. The input feature matrix $\mathbf{X}$ is projected into a latent representation via a batch-normalized MLP with ReLU activation:

$$\mathbf{H}_0 = \mathrm{ReLU}\big(\mathrm{BN}_0(\mathbf{W}_0\mathbf{X} + \mathbf{b}_0)\big). \tag{2}$$

Equation (2) defines the initial node embeddings consumed by the graph layers.

Graph Attention (GAT) layers. We apply two stacked graph attention layers; each layer aggregates neighbor information using attention weights and is followed by batch normalization and ELU activation:

$$\mathbf{H}_1 = \mathrm{ELU}\big(\mathrm{BN}_1(\mathrm{GAT}_1(\mathbf{H}_0))\big), \tag{3}$$

$$\mathbf{H}_2 = \mathrm{ELU}\big(\mathrm{BN}_2(\mathrm{GAT}_2(\mathbf{H}_1))\big). \tag{4}$$

Equations (3)–(4) capture successive neighborhood aggregation.

Classification head. The final node embeddings are passed to an MLP classifier and softmax to obtain per-node class probabilities:

$$\hat{\mathbf{Y}} = \mathrm{Softmax}\big(\mathrm{MLP}_{\mathrm{cls}}(\mathbf{H}_2)\big). \tag{5}$$

Equation (5) yields the predicted probability vector for each node.

3.5. Loss and optimization

To handle class imbalance, we use a class-weighted categorical cross-entropy:

$$\mathcal{L} = -\sum_{i \in \mathcal{T}} w_{y_i} \log\big(\hat{y}_{i,y_i}\big), \tag{6}$$

where $\mathcal{T}$ is the training set, w_{y_i} are class weights (inverse to class frequency), and $\hat{y}_{i,y_i}$ is the predicted probability of the true label. We optimize with Adam and apply weight decay, dropout, and early stopping.

Figure 1 shows the PS-GNN pipeline including feature projection, graph construction, attention-based message passing, and the classification head.

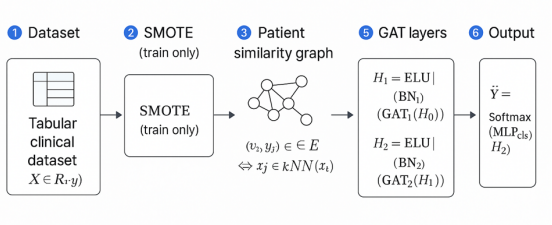

Figure 1: Overview of the Patient Similarity Graph Neural Network (PS-GNN) architecture: MLP projection, GAT-based message passing, and classification head.

4. Results and Discussion

4.1. Datasets and Evaluation Metrics

The proposed PS-GNN was evaluated on three benchmark datasets: *UCI Heart Disease* [8], *Cleveland Clinic* [9], and the *Framingham Heart Study* [10]. The UCI and Cleveland datasets contain five classes with diagnosis label $\texttt{num} \in \{0, 1, 2, 3, 4\}$, where 0 denotes no disease and 1–4 indicate presence with increasing angiographic severity (commonly defined by $> 50\%$ vessel narrowing). The Framingham dataset is binary classification $\in \{0, 1\}$.

TABLE 1: Summary of Datasets Used in the Study

Dataset	Samples	Features	Classes
UCI Heart Disease	920	13	5
Cleveland Clinic	303	13	5
Framingham Study	3658	15	2

Metrics. Model performance was measured using accuracy, macro-averaged precision, recall, and F1, computed as follows:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} 1\{y_i = \hat{y}_i\} \tag{8}$$

$$\mathrm{Prec}_{macro} = \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{9}$$

$$\mathrm{Rec}_{macro} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{10}$$

$$\mathrm{F1}_{macro} = \frac{1}{K}\sum_{k=1}^{K}\frac{2\,\mathrm{TP}_k}{2\,\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{11}$$

4.2. Performance Comparison with Baselines

PS-GNN was compared against Decision Tree (DT), Transformer-based, and Graph Attention Network (GAT) baselines. Table 2 presents the results across datasets.

PS-GNN either matched or outperformed the strongest baselines. On the **UCI Heart Disease** dataset, PS-GNN achieved 0.82 accuracy—equal to GAT but with slightly higher precision, recall, and F1. On **Cleveland Clinic**, PS-GNN attained the highest accuracy (0.85), improving upon GAT while maintaining balanced metrics. The largest margin appeared in the **Framingham Study**, where PS-GNN achieved 0.85 accuracy and F1 = 0.85, surpassing all baselines.

These improvements demonstrate the strength of combining MLP-based local feature encoding with graph-attention relational learning.

TABLE 2: Performance Comparison of Models Across Datasets (bold indicates best)

Model	Accuracy	Precision	Recall	F1 Score
UCI Heart Disease				
Decision Tree	0.73	0.72	0.73	0.72
Transformer	0.78	0.78	0.78	0.77
GAT	**0.82**	0.81	**0.82**	0.81
PS-GNN	**0.82**	**0.82**	**0.82**	**0.82**
Cleveland Clinic				
Decision Tree	0.70	0.71	0.70	0.71
Transformer	0.71	0.72	0.71	0.68
GAT	0.84	**0.84**	0.84	0.83
PS-GNN	**0.85**	**0.84**	**0.85**	**0.85**
Framingham Study				
Decision Tree	0.78	0.79	0.78	0.78
Transformer	0.70	0.71	0.70	0.70
GAT	0.74	0.75	0.74	0.74
PS-GNN	**0.85**	**0.86**	**0.85**	**0.85**

4.3. Receiver Operating Characteristic (ROC) Analysis

PS-GNN attained AUC ¿ 0.85 on UCI, indicating strong discriminative capability and reduced false classifications in clinical use.

On Cleveland data, PS-GNN preserved high sensitivity while lowering false positives, crucial for early intervention.

Across datasets, the ROC results confirm PS-GNN's robustness and reliable separation of CVD and non-CVD patients.

4.4. Confusion Matrix Analysis

On UCI, PS-GNN minimized both false positives and false negatives, reflecting balanced clinical prediction.

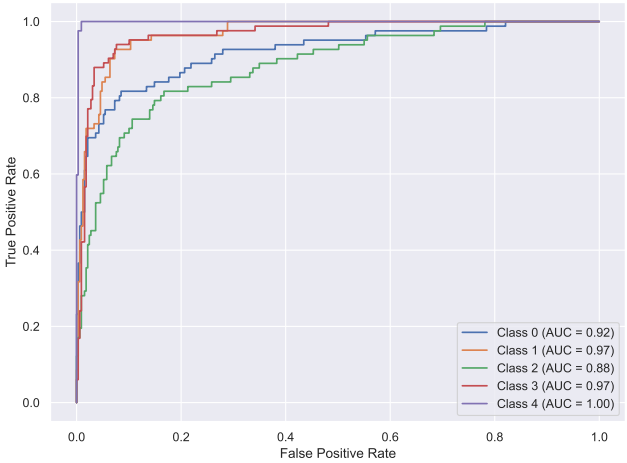

Figure 2: ROC Curve — UCI Heart Disease.

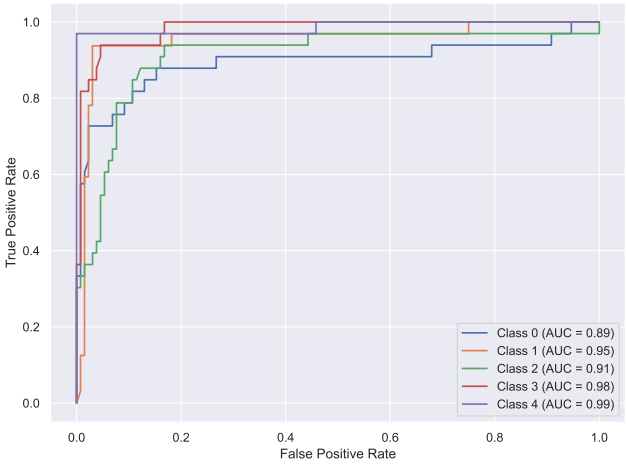

Figure 3: ROC Curve — Cleveland Clinic.

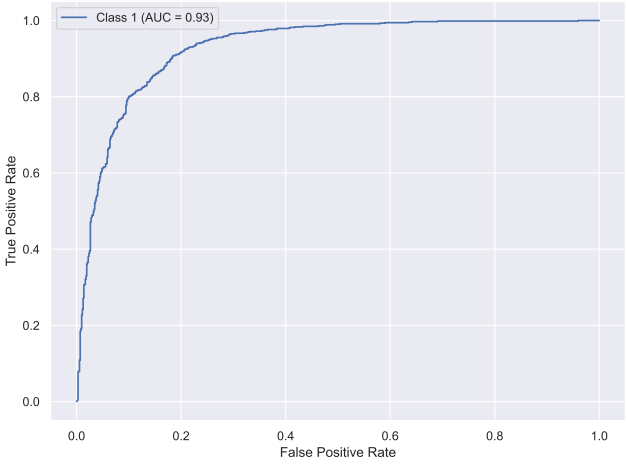

Figure 4: ROC Curve — Framingham Study.

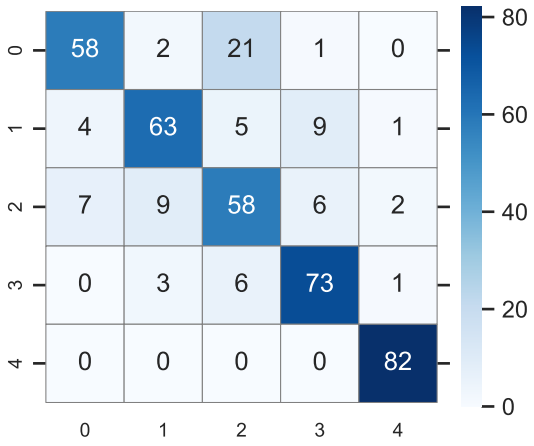

Figure 5: Confusion Matrix — UCI Heart Disease.

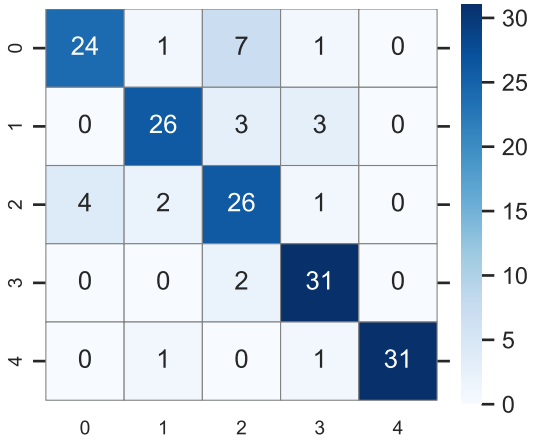

Figure 6: Confusion Matrix — Cleveland Clinic.

For Cleveland, the confusion matrix shows stable classification even with moderate class imbalance.

In the Framingham dataset, PS-GNN correctly identified most CVD cases, confirming its suitability for real-world screening.

4.5. Practical Implications

In a hospital setting, PS-GNN can be integrated into electronic health record (EHR) systems to automatically compare new patient profiles against historical data. If the model detects a high CVD-risk pattern, physicians can be alerted for timely intervention—potentially preventing heart attacks or strokes.

5. Conclusion

This paper proposed PS-GNN, a Patient Similarity Graph Neural Network combining MLP-based feature projections and attention-based graph aggregation for cardiovascular disease prediction. Evaluations on three benchmark datasets

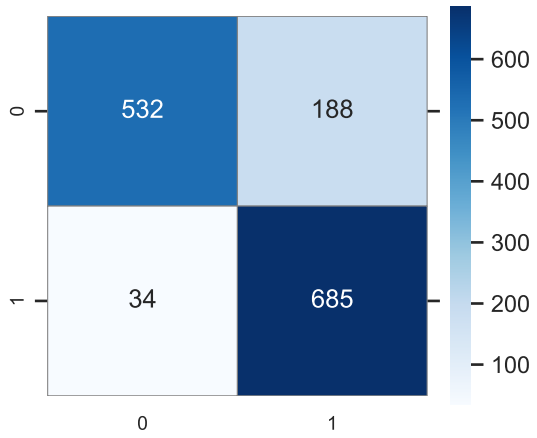

Figure 7: Confusion Matrix — Framingham Study.

show PS-GNN delivers robust, balanced performance, outperforming baseline methods in several key metrics. Future work will integrate explainability modules to support clinical interpretability and assess prospective performance in real-world EHR environments.

Acknowledgment

The authors thank the ICCAIT reviewers for constructive feedback that strengthened the manuscript.

References

[1] T. A. Gaziano, "Cardiovascular diseases worldwide," *Public Health Approach to Cardiovascular Disease Prevention and Management*, vol. 1, pp. 8–18, 2022.

[2] N. A. Mahoto, A. Shaikh, A. Sulaiman, M. S. Al Reshan, A. Rajab, and K. Rajab, "A machine learning-based data modeling for medical diagnosis," *Biomedical Signal Processing and Control*, vol. 81, p. 104481, 2023.

[3] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105400, 2020.

[4] E. Elsedimy, S. M. AboHashish, and F. Algarni, "A new cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23901–23928, 2024.

[5] R. R. Isnanto, I. Rashad, and C. E. Widodo, "Classification of heart disease using linear discriminant analysis algorithm," in *E3S Web of Conferences*, vol. 448, p. 02053, EDP Sciences, 2023.

[6] J. G. D. Ochoa and F. E. Mustafa, "Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses," *Artificial Intelligence in Medicine*, vol. 131, p. 102359, 2022.

[7] S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, and A. Moustafa, "A systematic review of graph neural networks in healthcare-based applications: Recent advances, trends, and future directions," *IEEE Access*, vol. 12, pp. 15145–15170, 2024.

[8] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease dataset," UCI Machine Learning Repository, 1989. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[9] Cleveland Clinic Foundation, “Cleveland heart disease dataset (via UCI Machine Learning Repository),” 1989. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease

[10] Framingham Heart Study, “Framingham Heart Study Dataset,” National Heart, Lung, and Blood Institute (NHLBI), Online dataset, 1948–present. [Online]. Available: https://www.framinghamheartstudy.org