Journal of Biomedical Informatics 178 (2026) 105015

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Interpretable Graph Convolutional Networks for cardiovascular disease risk prediction in patients with Type 2 Diabetes Mellitus

Ioannis Siachos [a,*], Maria Athanasiou [a], Konstantia Zarkogianni [a], Anastasia C. Thanopoulou [b], Konstantina S. Nikita [a]

[a] School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
[b] Medical School, University of Athens, Athens, Greece

ARTICLE INFO

Keywords:
Type 2 Diabetes Mellitus
Cardiovascular disease
Machine learning
Graph Neural Networks
Anomaly detection
Explainable artificial intelligence
Global surrogate model

ABSTRACT

Background : Cardiovascular disease (CVD) is the most prevalent complication of Type 2 Diabetes Mellitus (T2DM) and a leading cause of mortality in this population. Early and accurate CVD risk prediction is essential for timely intervention, yet traditional clinical risk calculators may overlook complex, non-linear relationships between risk factors, and have exhibited varying performance. Graph neural networks (GNNs) are able to capture these complex relationships but often lack interpretability.

Objective : The present study aims to develop and evaluate the first interpretable Graph Neural Network (GNN)-based framework for Cardiovascular Disease (CVD) risk prediction in patients with Type 2 Diabetes Mellitus (T2DM). We introduce a novel approach that integrates a GNN classifier with a rule-based surrogate model to generate clinically meaningful explanations for the model's predictions, addressing the critical need for both high accuracy and transparency in clinical AI.

Methods : A population graph of 560 T2DM patients was constructed using demographic, lifestyle, laboratory, and treatment data. A GNN-based classifier was trained by leveraging a loss function originally designed for graph-based anomaly detection to address class imbalance. Post-hoc interpretability was achieved through the deployment of a RuleFit surrogate model, combining decision tree ensembles and a sparse linear model to extract global, rule-based explanations.

Results : The proposed model achieved an AUC of 0.786±0.076, exceeding all benchmark methods and outperforming prior best-reported results on this dataset by over 7% in terms of the AUC, and produced well-calibrated probabilities (Brier score: 0.053±0.021). RuleFit explanations aligned with established CVD risk factors, while revealing intermediate-risk patterns, such as residual dyslipidemia despite treatment, that may warrant earlier intervention.

Conclusion : The proposed interpretable GNN framework demonstrated its ability to provide reliable CVD risk estimates while offering transparent, clinically relevant explanations. These findings support its potential integration into CVD risk screening tools for patients with T2DM, paving the way for real-world clinical implementation.

1. Introduction

Diabetes Mellitus (DM) is a chronic metabolic disease, characterized by increased blood glucose levels. Type 2 Diabetes Mellitus (T2DM) is the most common type of DM accounting for over 90% of all DM cases worldwide. In T2DM, hyperglycemia is the result of insulin resistance, that is the inability of the body cells to fully respond to insulin. T2DM is commonly diagnosed in older adults, yet the exact time of onset is usually impossible to determine due to the asymptomatic nature of the disease at its early stages. Prolonged elevated blood glucose levels may lead to the incidence of long-term, serious, and mortality-related complications including micro- and macro-vascular diseases. According to the International Diabetes Federation (IDF), the global prevalence of DM is high and rising across all regions, accounting for 589 million adults in 2024 globally, whereas this number is projected to rise to 853 million by 2050 [1]. Due to its association with acute episodes and complications, the disease caused 3.4 million deaths and at least 1 trillion dollars in health expenditure in 2024. T2DM, in particular, is currently the 8th leading cause of disease burden globally and estimated to

* Corresponding author.
E-mail address: isiachos@ubitech.eu (I. Siachos).

https://doi.org/10.1016/j.jbi.2026.105015
Received 23 September 2025; Received in revised form 2 March 2026; Accepted 11 March 2026
Available online 13 March 2026

become the second leading cause by 2050. Optimal management of T2DM requires glycemic control through effective lifestyle behavioral changes and appropriate medication – insulin or oral medication – if needed. Moreover, it is critically important to manage blood pressure and blood cholesterol levels along with performing regular screening for possible complications in order to initiate preventive measures.

Cardiovascular disease (CVD) constitutes the most prevalent complication of DM and the main cause of mortality in this population. The CVD risk in patients with T2DM is estimated to be 2- to 4- folds higher compared to the general population, thus, posing the need to develop evidence-based recommendations for the prevention and management of CVD in patients with DM [2]. In response to this need, the European Society of Cardiology (ESC) in collaboration with the European Association for the Study of Diabetes (EASD) has recently released relevant guidelines recommending CVD risk-based management while adopting CVD risk scores dedicated to the DM population [3].

Statement of Significance	
Problem	Cardiovascular disease (CVD) is the leading cause of morbidity and mortality in type 2 diabetes mellitus (T2DM). Current risk assessment methods often fail to capture risk factor interactions, thus limiting timely, personalized intervention.
What is Already Known	Graph neural networks (GNNs) can model complex, non-linear relationships in patient data, but their limited interpretability hinders clinical adoption. Their use in CVD risk prediction for T2DM patients has not been investigated.
What this Paper Adds	This work introduces an interpretable GNN framework, combining graph anomaly detection with clinically understandable rules, able to reveal hidden risk patterns even in challenging, imbalanced datasets. This approach bridges predictive performance with transparency, supporting earlier, tailored CVD prevention in T2DM.
Who Would Benefit from the Knowledge in this Paper	(i) Patients with T2DM who may gain access to earlier, more targeted CVD prevention, (ii) clinicians supported in making evidence-based decisions, (iii) researchers applying graph-based models in other disease contexts, (iv) policymakers seeking to integrate accurate, transparent AI tools into clinical workflows.

2. Related work

The most thoroughly studied CVD risk calculators [4], such as the Framingham, SCORE, and DECODE, have been based on data from the general population, thus underestimating the CVD risk in DM individuals. This has oriented many studies towards building CVD risk prediction models focusing on DM [5]. In this direction, survival analysis [6], linear regression [7], logistic regression [8], and hidden Markov models [9] have been utilized for the models' development, thus assuming the existence of linear relationships between the considered risk factors and the CVD risk while ignoring potential complex interactions between risk factors. Among the proposed models, the T2DM-specific UKPDS risk engine has demonstrated varying performance, primarily due to its data selection criteria, which have resulted in poor representation of a broader spectrum of T2DM patients, whereas the RECODe risk calculator, which has been based on data from the ACCORD study, has shown moderate-to-good discrimination and calibration in assessing the CVD risk in T2DM patients [10].

Attempts towards performance improvement in CVD risk prediction models for DM have been pursued through the investigation of more sophisticated approaches, founded on machine learning methods, including neural networks [11], support vector machines [12], self-organizing maps [13], as well as tree-based methods such as decision trees [14], and, most recently, the extreme Gradient Boosting (XGBoost) algorithm [15]. Moreover, the ability of deep learning models to effectively capture non-linearities in DM patient datasets in the case of CVD has also been explored [16]. All of the previous approaches have focused on modeling the complex interactions between risk factors, usually represented in the Euclidean space, while ignoring the existence of relationships between different patient profiles. Graphs offer a powerful and intuitive way to model individuals (nodes) and the relationships or similarities (edges) between individuals. In [17] it has been shown that creating relatively sparse geometric graphs from the tabular data of a dataset, even in the absence of information of relationships between the data points, provides an effective tool for modeling hidden relationships between features and patients, which can increase classification performance when fed to a Graph Convolutional Network (GCN) [18].

Graph Neural Networks (GNNs) constitute a class of artificial neural networks, able to unveil the complex relationships that can be found in non-Euclidean domains represented in the form of a graph. GNNs have been recently applied to the biomedical informatics domain, yielding promising results in prediction of pathogenicity in multi-type variants [19], disease diagnosis and prediction [20,21], neuroscience [22,23], survival analysis [24,25], source identification of infectious diseases [26,27], patient similarity learning [28] and side effects prediction [29]. Representing populations as a graph and transforming the disease prediction problem into a node classification task has been explored in [30–32].

To ensure the reliability and trustworthiness of risk prediction models in health and facilitate their adoption in clinical decision making, research has recently emphasized on harnessing inherently interpretable models as well as exploring various post-hoc model-agnostic interpretability techniques towards the generation of explanations on the black-box models' decisions. Along these lines, different interpretable risk prediction models have been proposed, aiming at enhancing human understanding on the models' reasoning process, increasing transparency, and ensuring user trust. Apart from inherently interpretable CVD risk prediction models, which have been based on the use of linear regression models and decision trees, a variety of interpretable approaches, leveraging different post-hoc interpretability techniques, have been proposed for estimating the CVD risk incidence. Global surrogate models, Partial Dependence Plots, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP) constitute some of the most widely investigated interpretability techniques that have been used to provide insights about features' contribution in the models' decision process. Furthermore, recent research in interpretable GNNs is advancing beyond basic attention mechanisms, with novel frameworks including the Graph Multilinear neT (GMT) [33] and the Tree-like Interpretable Framework (TIF) [34], aiming for more faithful and multi-granular explanations of GNN predictions. Concurrently, the integration of GNNs with Healthcare Knowledge Graphs [35] and Large Language Models [36] is enhancing their reasoning capabilities and the transparency of their applications in complex biomedical domains. However, a limited number of interpretable models focusing on the assessment of the CVD risk in DM patients can be identified in the literature.

In the present study, an interpretable approach, based on the combined use of Graph Neural Networks (GNNs) and the RuleFit algorithm [37], is proposed towards the development of an interpretable personalized risk prediction model for the fatal or non-fatal CVD incidence in T2DM. The ability of GNNs to model latent relationships between patients through a comprehensive weighted patient network and address the graph anomaly detection problem in the case of class imbalance is leveraged towards exploring nonlinear complex interactions between patients and CVD-related risk factors and effectively handling the unbalanced nature of the used dataset. A global hybrid surrogate model, combining the predictive power of tree-based ensembles with the interpretability of linear models, is deployed in order to provide meaningful rule-based explanations on the GNN model's outputs. To the best of the authors' knowledge, this is the first work proposing an interpretable GNN-based prediction model for the calculation of the CVD risk in T2DM.

Table 1
Demographic characteristics of the study cohort (N=560).

Characteristic	Value
Age (years), mean ± SD	58.56 ± 10.70
Sex, n (%)	
Male	263 (46.96%)
Female	297 (53.04%)
Ethnicity, n (%)	
Greek	560 (100%)

Table 2
Baseline prevalence of comorbidities in the study cohort ($N = 560$).

Comorbidity	n (%)
Dyslipidemia	208 (37.1)
Hypertension	80 (14.3)
Retinopathy	85 (15.2)
Microalbuminuria	60 (10.7)
Macroalbuminuria	33 (5.9)
Chronic Kidney Disease	9 (1.6)
Peripheral Neuropathy	42 (7.5)
Autonomic Neuropathy	3 (0.5)
Angina	14 (2.5)
Myocardial Infarction	7 (1.3)
Hemorrhagic Stroke	4 (0.7)

Table 3
Continuous variables of the dataset.

Variable	Average ± Standard deviation
Diabetes duration	7.67 ± 7.37 (years)
Body Mass Index (BMI)	29.49 ± 5.54
Pulse Pressure	56.75 ± 15.80 (mmHg)
Glycosylated Hemoglobin	7.43 ± 1.81 (%)
Fasting Glucose	165.15 ± 56.15 (mg/dL)
Total Cholesterol	226.64 ± 50.04 (mg/dL)
Triglycerides	167.39 ± 110.81 (mg/dL)
HDL Cholesterol	48.35 ± 16.46 (mg/dL)

Table 4
Categorical variables of the dataset.

Variable	Number of patients
Smoking Habit:	
Non Smokers	289 (51.61%)
Current Smokers	146 (26.07%)
Ex-smokers	125 (22.32%)
Parental History of Diabetes:	
No	304 (54.28%)
Yes	256 (45.72%)
Lipid-lowering therapy:	
No	469 (83.75%)
Statins	74 (13.21%)
Fibrates	17 (3.04%)
Aspirin:	
No	509 (90.89%)
100 mg	44 (7.85%)
325 mg	7 (3.03%)

3. Materials and methods

3.1. Data

Data collected from the 5-year follow-up of 560 T2DM patients at the Hippokration General Hospital of Athens (1996–2007) were used for the model's development and evaluation. The study cohort includes patients with a confirmed T2DM diagnosis and complete baseline records for the variables under study. The dataset is characterized by a severe class imbalance, which presents a significant modeling challenge. Specifically, a small fraction of the patients $(41, 7.32\%)$ presented a CVD incident during their follow-up. Four patients experienced stroke and the rest experienced coronary heart disease (CHD). Baseline demographic, lifestyle, laboratory, and treatment data that provide adequate information on a T2DM patient profile were used to compose the model's feature space [13]. The considered variables represent the full set of routinely monitored parameters for T2DM management and CVD risk assessment, which were recorded by clinicians in accordance with current guidelines and established literature. To support model interpretability and facilitate clinical translation, this compact set of clinically relevant and consistently collected variables was directly utilized without any prior exploratory or automated feature selection procedure. An overview of the continuous and categorical dataset variables – none of which contained missing values – is presented in Tables 3 and 4. The demographic characteristics and key comorbidities of the study cohort are summarized in Tables 1 and 2.

3.2. Methods

3.2.1. Conceptual framework

The complexity of interactions between the multitude of factors, underpinning the manifestation of T2DM and its complications, motivated the deployment of graphs as a means for the efficient representation and identification of complex patterns present in the data. Moreover, the inherent highly expressive capability of GNNs in learning graph representations through message passing was leveraged towards building the proposed CVD risk prediction model.

To address the imbalanced nature of the dataset, an approach based on graph anomaly detection was adopted, building on GNNs' ability to detect anomalous patterns in graphs. Handling class imbalance through anomaly detection algorithms [38,39] involves solving a classification problem, where data belonging to the minority class are considered anomalous, and the rest normal. In the case of graphs, algorithms addressing anomaly detection usually follow an unsupervised approach, but may also consider label information, referring to the characterization of observations either as normal or anomalous [40]. GNNs have been adopted to efficiently and intuitively detect anomalies by examining graph topology/structure and node attributes/instance information simultaneously to extract anomalous patterns from graphs, even those with highly complex structures or attributes.

Within the framework of the present study, a GNN-based model with a loss function that was originally proposed for an anomaly detection task, accepting data from the normal and abnormal class and learning to separate them [41], was deployed. Unlike standard cost sensitive learning methods, including weighted cross-entropy or focal loss, which address imbalance by increasing the penalty for misclassifying minority samples, the adopted approach reframed the problem entirely. The chosen hypersphere-based loss function was designed to learn a geometrically compact representation for the 'normal' (Non-CVD) patient majority in the embedding space, while encouraging embeddings of high-risk (CVD) patients to lie far from this 'normal' center, treating them as structural anomalies within the graph. This method is particularly well-suited for GNNs, as it leverages both patient features and the learned inter-patient relationships to achieve more robust class separation than simple loss re-weighting. In this context, the GNN model operated in a transductive learning setting, where only training nodes contributed labels to the loss, but the model had access to the entire graph structure – including nodes with withheld outcomes (i.e., validation and test patients) – allowing it to exploit relational information from these 'unlabeled' nodes during training. The resulting anomaly scores were transformed into the corresponding probabilities via an exponential cumulative distribution function.

To facilitate user trust and enhance the reliability and transparency of the model's decision process, a post-hoc interpretability approach was deployed, combining decision tree ensembles and a sparse linear

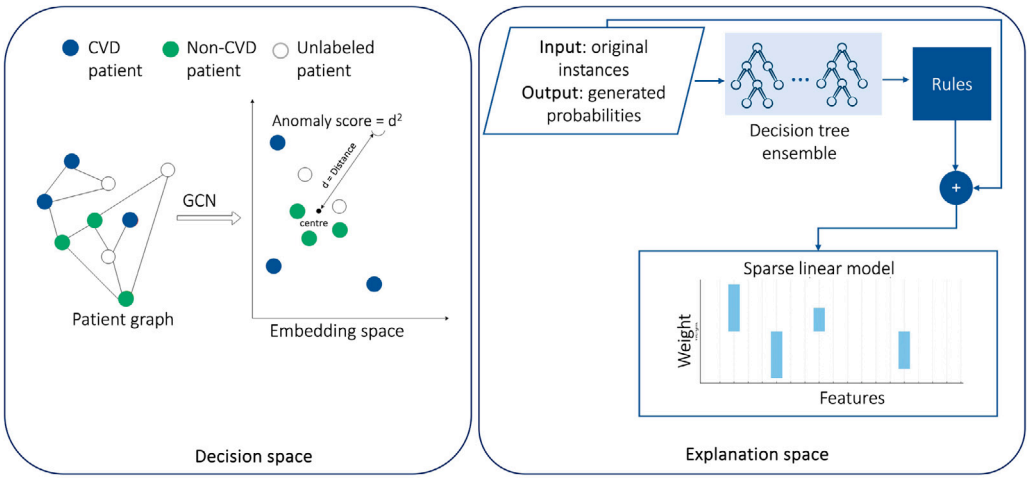

Fig. 1. Conceptual framework of the proposed model. It comprises the decision space, involving the GNN-based prediction model, and the explanation space, featuring a decision-tree-based global surrogate model and a sparse linear model for the generation of rule-based explanations. In this conceptual depiction, 'Unlabeled patient' refers to a node whose CVD outcome is withheld during model training (i.e., a member of the validation or test set), but whose features and graph connections are still leveraged by the GNN in its transductive learning process.

model towards building a global surrogate model able to generate rule-based explanations [42]. An overview of the proposed model's conceptual framework is depicted in Fig. 1.

3.2.2. Population graph construction

Inspired by [31,32], a population graph was constructed in which each patient was represented by a node linked to the patient's corresponding feature vector. Edges between nodes were added according to the value of the following distance function:

$$D(i,j) = exp(\frac{\left\|\boldsymbol{x}_i(S_2) - \boldsymbol{x}_j(S_2)\right\|^2}{2\sigma^2}) * \frac{1}{|S_1|} * \sum_{f \in S_1} \delta(\boldsymbol{x}_i(f), \boldsymbol{x}_j(f)) \quad (1)$$

where i, j are the indices of two T2DM patients, S_1 and S_2 represent the sets of the considered categorical and continuous features, respectively, $\boldsymbol{x}_i(S_2)$ is a vector containing the continuous features of patient i, $\boldsymbol{x}_i(f)$ is the value of feature f for patient i, $\delta(x, y)$ is a function that returns 1 if $x = y$, σ is equal to the mean value of $\left\|\boldsymbol{x}_i(S_2) - \boldsymbol{x}_j(S_2)\right\|^2$ over all pairs of patient indices i, and j and $\boldsymbol{x}_i(B)$ are normalized such that $\left\|\boldsymbol{x}_i(S_2)\right\|_1 = 1$.

The values of the distance function D were calculated for all node pairs u and v with corresponding indices i and j and stored in a $560x560$ matrix $\boldsymbol{A}$ for which $A_{ij} = D(i,j)$. Then, according to a predefined threshold γ, which was optimally tuned, the values of the matrix $\boldsymbol{A}$ below γ were set to 0 and the rest to 1. The matrix $\boldsymbol{A}$ represented the adjacency matrix of the population graph. Each node u was assigned a label y_u equal to one or zero depending on whether the corresponding patient had developed or not fatal or non-fatal CVD during their 5-year follow-up period. Following this procedure, the CVD risk prediction problem was formulated as a node classification task.

3.2.3. GNN model

The proposed GNN model was based on a GNN architecture introduced in [41], which employs GCNs to compute node embeddings taking into account both local node features and their K-hop neighborhood in the graph. This representation learning framework is designed to facilitate the detection of anomalous nodes by mapping normal nodes to embeddings that lie close to a center vector $\boldsymbol{c}$, while anomalous nodes are encouraged to deviate significantly from this center. To address the imbalanced nature of anomaly detection tasks, a differential approximation of the AUC was utilized, which has been shown to be effective in imbalanced settings. A comprehensive description of the original architecture and its integration into the proposed approach, adapted to the specific requirements of this study, is provided in the remainder of this subsection.

Let $G = (V, \boldsymbol{A}, \boldsymbol{X})$ represent the patient graph, where V is the set of N=560 nodes, $\boldsymbol{A} \in \mathbb{R}^{(NxN)}$ is the graph adjacency matrix, and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{(NxD)}$ is the node feature matrix with D denoting the number of features. Let $Norm$ and $AbNorm$ represent the index set of patients without and with CVD, corresponding to class 0 (normal) and class 1 (abnormal), respectively. The model was trained to assign an anomaly score to all nodes, with the objective function being optimized on the labeled training nodes. For unlabeled nodes within the graph (i.e., those belonging to the validation and test sets), the model's generalization ability was evaluated by how well it assigned low scores to normal (class 0) and high scores to abnormal (class 1) patients.

The model architecture was based on a two-layer GCN, which computed node embeddings by aggregating information from local neighborhoods up to K hops based on the following equation:

$$\boldsymbol{H} = \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{X} \boldsymbol{W}^{(0)}) \boldsymbol{W}^{(1)} \quad (2)$$

where $\tilde{\boldsymbol{A}}$ is the graph adjacency matrix with the addition of self loops ($\tilde{A_{ij}} = 1, \forall i \in [1, 2, \ldots, N]$), $\tilde{\boldsymbol{D}}$ is a diagonal matrix where each $\tilde{D_{ij}}$ contains the number of neighbors of node i, σ is a non-linear activation function, $\boldsymbol{X}$ is the feature matrix with the continuous and the one-hot encoded categorical values, $\boldsymbol{W}^{(0)}$ is the weight matrix of the first layer of the GCN, $\boldsymbol{W}^{(1)}$ is the weight matrix of the second layer of the GCN, and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]^T \in \mathbb{R}^{(Nxk)}$ is the matrix that contains the final embedding, $\boldsymbol{h}_i$,for each node i. During training, the Dropout regularization method was utilized [43] and σ was set to be the GELU activation function [44].

For each node u_i, the model produced an anomaly score $a(u_i)$ computed as the squared Euclidean distance between its embedding $\boldsymbol{h}_i$ and the center $\boldsymbol{c}$:

$$a(u_i) = \left\|\boldsymbol{h}_i - \boldsymbol{c}\right\|^2 \quad (3)$$

The center vector $\boldsymbol{c}$ was calculated as the mean of embeddings for the normal-class nodes $\boldsymbol{h}_i, i \in Norm$ of the training set in the initial forward pass with randomly initialized weights, reflecting the assumption that embeddings of normal nodes should lie near $\boldsymbol{c}$, while abnormal ones should lie farther away. This derivation of $\boldsymbol{c}$ solely from the distribution of normal patients in the training set prevented the risk of distribution leakage, ensuring that the representation of 'normality' was defined solely by data available during training. Thus, the anomaly scores assigned to test patients quantified deviations from this training-defined embedding space rather than from statistics influenced by unseen test

data. Moreover, while the center vector is considered to be the center of a hypersphere where each normal node embedding lies inside its radius and each abnormal one lies outside it, its calculation was not part of the original model training procedure and, thus, no explicit decision threshold for classifying an instance as normal or anomalous was learned.

The model was trained by minimizing a custom loss function incorporating both hypersphere compactness for normal nodes and discriminative separation from abnormal nodes:

$$L(\theta) = L_{nor}(\theta) - \lambda L_{AUC}(\theta) \tag{4}$$

where $\theta = \{\boldsymbol{W}^{(0)}, \boldsymbol{W}^{(1)}\}$ denotes the learnable parameters and λ is a hyperparameter controlling the trade-off between objectives. The first term of the loss function enforces compactness of normal node embeddings around the center by minimizing the volume of the hypersphere that encloses normal nodes:

$$L_{nor}(\theta) = \sum_{u \in Norm} \frac{\|\boldsymbol{h_u} - \boldsymbol{c}\|^2}{|Norm|} \tag{5}$$

The second term is a differential approximation of the AUC, encouraging higher anomaly scores for abnormal nodes:

$$L_{AUC}(\theta) = \frac{1}{|Norm||AbNorm|} \sum_{i \in AbNorm} \sum_{j \in Norm} f(a(u_i) - a(u_j)) \tag{6}$$

where f is the sigmoid function. Model optimization was performed using the Adam optimizer [45].

To enable probabilistic interpretation of anomaly scores, each score $a(u_i)$ was transformed into a probability estimate indicating the likelihood that patient u_i belonged to class 1 (CVD). This was achieved by computing a scaling parameter μ, defined as the inverse of the mean anomaly score across the training set, and applying the exponential cumulative distribution function:

$$F(x;\mu) = \begin{cases} 1 - exp(-\mu * x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{7}$$

Under this formulation, a score of zero mapped to a probability of 0, a score equal to the mean anomaly score yielded a probability of 0.5, and the probability asymptotically approached 1 as the score tended towards infinity. This transformation enabled the model to produce interpretable probability estimates consistent with the underlying anomaly score distribution.

3.2.4. Interpretability

To derive insightful explanations on the model's decisions, an interpretable surrogate model was employed based on the use of the RuleFit algorithm [37]. RuleFit constructs a set of human-interpretable rules by extracting decision paths from an ensemble of decision trees, and subsequently fits a sparse linear model using both the original input features and the derived rules as predictors. A key advantage of this approach lies in its inherent interpretability as the surrogate model can yield transparent and informative explanations regarding the decision boundaries of the original, more complex predictive model. For a given instance, interpretability is achieved by identifying the subset of rules that the instance satisfies. Each rule is associated with a coefficient in the linear model, indicating its contribution to the overall prediction. The magnitude of the coefficient reflects the importance (i.e., weight) of the rule, while the sign denotes the direction of its influence with positive (negative) coefficients supporting classification into the positive (negative) class.

4. Results

4.1. Evaluation framework, hyperparameters' tuning, and model training

To ensure robust hyperparameter optimization and unbiased generalization performance estimates, a 10 × 4 nested stratified cross-validation scheme was employed [46], with inner loops used for tuning and outer loops for model evaluation, preserving class imbalance across folds [47]. Model performance was assessed using appropriate metrics for imbalanced classification tasks, including the Area Under the ROC Curve (AUC), F1-score, Precision, and Recall. Brier Score was deployed to evaluate probability calibration, quantifying the mean squared difference between predicted and true outcome probabilities, with lower values indicating better calibration [48].

Table 5
Search space of the hyperparameter tuning procedure.

Hyperparameter	Values
Hidden Layer Neurons	4,8,16,32,64,128
γ	0.3, 0.4, 0.5, 0.6, 0.7
Output Layer Neurons	4, 8, 16, 32, 64, 128
Dropout	0, 0.1, 0.2, 0.3, 0.5, 0.6
Adam Learning Rate	1e−2, 5e−3, 1e−3, 1e−4, 5e−4
λ	$0, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6$
Adam Weight Decay	0, 1e−2, 5e−2, 5e−3, 1e−3, 5e−4, 1e−4, 5e−5, 1e−6

Hyperparameter optimization was performed by applying grid search over a predefined hyperparameter space. The search space, presented in Table 5, featured the number of neurons in the hidden and output layers, the adjacency matrix threshold γ, the training loss hyperparameter λ, the dropout rate, the Adam learning rate, and Adam weight decay. The optimal configuration was selected based on the obtained average AUC score on the inner loop of the nested cross-validation scheme.

During model training, each iteration of the outer cross-validation loop involved partitioning the data into training and validation sets using a 90:10 split. The model was trained for up to 2000 epochs, with early stopping applied based on the F1-score obtained on the validation set. The scaling parameter μ of Eq. (7) was calculated by considering the anomaly score on the validation set of the outer loop. The optimal decision threshold was selected within the range [0.1, 0.9] based on the obtained F1-score on the validation set.

4.2. Evaluation of models' performance

The results obtained from the applied cross-validation scheme are summarized in Table 6. The model yielded satisfactory discrimination performance, achieving an AUC score of 0.786 ± 0.076, and demonstrated its ability to handle the unbalanced nature of the data, as reflected by a Precision of 0.711 ± 0.210, Recall of 0.630 ± 0.130, and F1-score of 0.641 ± 0.100. In terms of calibration, the model produced well-calibrated probability estimates, as indicated by the obtained Brier Score of 0.053 ± 0.021.

To validate the model's underlying learning mechanism, we analyzed the distribution of the final anomaly scores (i.e., the squared Euclidean distance from the center c) across all held-out test sets. The results confirmed that the hypersphere-based loss function successfully learned a discriminative embedding space. Patients in the normal (non-CVD) group had a mean anomaly score of 1.15 ± 0.65, reflecting their intended compactness around the normative center. In contrast, patients who developed CVD (abnormal group) were mapped significantly farther away, with a mean anomaly score of 3.20 ± 1.85. An independent samples t-test performed on these scores confirmed that this difference was statistically significant ($p < 0.001$), providing direct evidence that our anomaly detection framework effectively separates the two classes.

4.3. Comparison with the state of the art

The dataset used in the present study has been previously explored in [13,15], where various machine learning algorithms were applied for CVD risk prediction in T2DM. Compared to these approaches, the proposed model achieved superior performance, exhibiting an AUC improvement of over 7% with respect to the best results reported

Table 6
Performance of the proposed GNN-based model and comparative assessment with the previously proposed XGBoost-based model within the applied cross-validation scheme (metrics reported as average ± standard deviation across folds).

Metric	Proposed GNN-based model	XGBoost-based model
AUC	**0.786 ± 0.076**	0.724 ± 0.060
F1-score	**0.641 ± 0.100**	0.367 ± 0.120
Precision	**0.711 ± 0.210**	0.353 ± 0.230
Recall	**0.630 ± 0.130**	0.515 ± 0.050
Brier Score	**0.053 ± 0.021**	0.141 ± 0.030

in the aforementioned studies. This performance gap indicated the proposed model's ability to handle the class imbalance and limited sample size inherent in the dataset more effectively. The proposed approach also outperformed baseline regression models currently used in clinical practice, including the UKPDS Risk Engine and Bayesian Logistic Regression (BLR) [13]. This provided further evidence regarding GNNs' ability to capture complex relational structures within the data more effectively than traditional models, suggesting that GNN-based approaches may offer a more suitable framework for CVD risk prediction in T2DM patients.

Furthermore, to provide a direct performance benchmark, the XGBoost classifier, which was recently applied to the same prediction task [15], was comparatively assessed against the proposed model, given its relatively low computational complexity and fast training time. To this end, XGBoost was trained using the same nested cross-validation scheme and fold structure as the proposed GNN model. A random search over the positive class weight parameter in the range [0, 1] was conducted, along with a threshold-moving technique on the validation set, mirroring the approach used for the GNN model. As depicted in Table 6, the proposed GNN model demonstrated significantly stronger performance across all metrics. Compared to the XGBoost classifier, our model achieved an 8.6% higher AUC (0.786 vs. 0.724), a 75% higher F1-score (0.641 vs. 0.367), and a 101% higher precision (0.711 vs. 0.353). This highlights the GNN's superior ability to effectively handle the severe class imbalance and complex relationships within this dataset.

4.4. Interpretability

4.4.1. Surrogate model performance

The RuleFit algorithm was implemented using the skope-rules Python library, a well-established package for rule extraction. To ensure a standardized and reproducible approach, we utilized the library's default hyperparameter settings, which are optimized for general-purpose rule induction and allowed us to focus the evaluation on the GNN's learned representations rather than on tuning the surrogate model itself.

The RuleFit surrogate model was trained on the same dataset with each sample labeled with the predictions of the original GNN model. To evaluate the surrogate's ability to approximate the behavior of the original model, the R-squared measure was employed. For each train-test split of the outer cross-validation loop, the surrogate model was fitted on the corresponding training set. To address the class imbalance present in the data, a random 40% undersampling was applied to the negative class during training. The surrogate model achieved an average R-squared score of 92.51%, indicating a strong correspondence with the GNN's predictions. This high level of agreement suggested that the RuleFit model served as a faithful and interpretable approximation of the original GNN model, enabling the extraction of meaningful rules that explain model behavior for individual instances.

4.4.2. Rule-based explanations

The analysis of the rule-based explanations generated by the surrogate RuleFit model provided useful insights into the factors driving the proposed model's predictions and enabled an assessment of their alignment with established clinical knowledge and recognized cardiovascular risk factors. Fig. 2 presents the most influential rules – determined by their corresponding coefficients in the RuleFit surrogate model – that were satisfied by four representative patient cases from the test set of the first cross-validation fold, corresponding to a True Positive (TP), a True Negative (TN), a False Positive (FP), and a False Negative (FN) prediction.

Rules linked to the TP prediction (Fig. 2(a)) highlighted established cardiovascular risk factors, including elevated total cholesterol, advanced age, and low HDL levels, especially in patients undergoing lipid-lowering therapy. Conversely, the identified rules associated with the TN prediction (Fig. 2(b)) predominantly reflected feature combinations characteristic of low-risk patient profiles, including shorter diabetes duration, moderate triglyceride levels, absence of lipid-lowering therapy, adequate HDL levels, and controlled blood pressure. These findings suggest that the proposed model correctly identified protective factors against CVD and exhibited sensitivity to clinically meaningful indicators of CVD risk, aligning with clinical knowledge.

In the FP prediction case (Fig. 2(c)), the highlighted rules reflected the model's sensitivity to individual risk factors, such as very high total cholesterol, that triggered a positive prediction despite the absence of an actual event. Overall, this case indicated overestimation of the CVD risk due to the increased influence of the patient's lipid profile as well as possibly uncaptured protective factors or individual patient variability not represented in the training data. For the FN prediction (Fig. 2(d)), the identified rules involved specific value ranges for various features (e.g., triglycerides, blood pressure, and HbA1c) that the model may have interpreted as non-critical, thus leading to an underestimation of risk. Moreover, although known risk factors such as age or smoking habit were identified, these may have been outweighed by the influence of protective features, guiding the model towards the calculation of lower risk scores.

5. Discussion

The present study demonstrated that GNN-based models can provide reliable predictions for CVD risk stratification in patients with T2DM, even in the presence of class imbalance and limited sample sizes. To achieve this, the proposed approach leveraged individual-level attributes and inter-patient relationships modeled in a graph representation to capture higher-order dependencies present in the data. Moreover, it adapted a graph anomaly detection method, originally developed for the unsupervised identification of anomalous patterns, to the supervised prediction of CVD risk in a severely imbalanced setting, enhancing the model's ability to detect minority-class (high-risk) cases while ensuring balanced classification performance.

During model development, conventional cost-sensitive learning approaches, including weighted cross-entropy and focal loss, were initially investigated for model training. In both cases, the GNN failed to learn meaningful discriminatory patterns; specifically, the model converged to predicting exclusively the majority class, labeling all nodes as 'Normal' and yielding an average AUC of 0.50. These results demonstrated that simply re-weighting the classification loss was insufficient for capturing the subtle minority-class signal present in this severely imbalanced clinical dataset.

Data-level imbalance handling methods such as SMOTE [42] were also explored but did not improve discriminative performance. These experiments resulted in an average AUC of 0.50, as the model assigned all validation nodes to the majority ('Normal') class. This outcome is expected in small, high-dimensional clinical datasets, where synthetic minority samples often fail to approximate the true underlying distribution and may introduce additional noise.

Along these lines, the conceptual reframing of the task, from classification to anomaly detection, emerged as a key element of the proposed

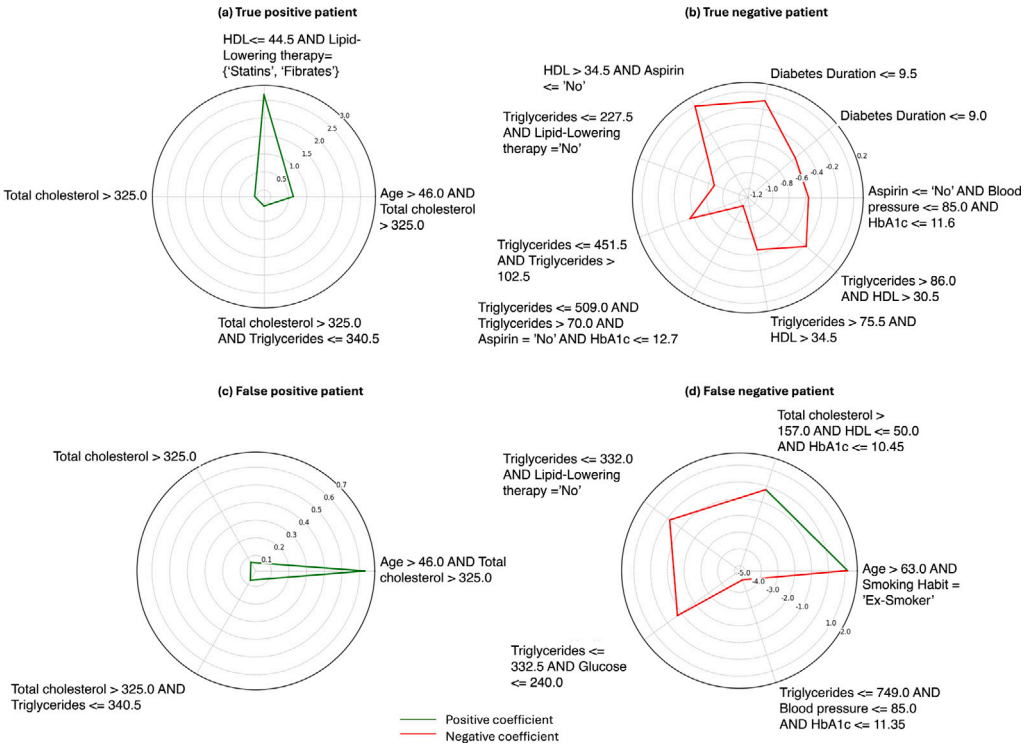

Fig. 2. Rules satisfied by a true positive (a), a true negative (b), a false positive (c), and a false negative (d) patient case. For each patient case, the subset of rules satisfied was identified along with the corresponding coefficients, indicating both the direction and magnitude of their contribution to the prediction.

model architecture, that was essential for addressing the aforementioned limitations of conventional approaches. The training objective was to learn a geometrically compact, normative representation of the majority (non-CVD) class within the embedding space, with the hypersphere-based loss function serving as the direct mechanism for enforcing this structure by treating high-risk patients as topological anomalies distant from the normative center. This formulation is particularly synergistic with the GNN architecture: by combining node features with relational information from neighboring patients, the model learns a more robust and holistic separation between high-risk and low-risk individuals than can typically be achieved through simple instance re-weighting. The integrated interaction between the graph structure, anomaly-driven objective, and hypersphere-based separation is therefore central to the model's ability to discriminate rare CVD cases in the context of severe class imbalance.

Using a nested cross-validation evaluation, the proposed framework yielded satisfactory discrimination ability, reflected in an AUC score of 0.786 ± 0.076, and was effective in identifying high-risk patients despite the severe class imbalance, while maintaining a low false positive rate, as indicated by the obtained precision, recall, and F1-score. Furthermore, the model produced well-calibrated probability estimates, ensuring that its outputs corresponded well to the actual observed risks.

Comparison with the state of the art showed that the proposed model outperformed all benchmark methods, including the best-performing models reported in prior work on this dataset [13,15] by a margin exceeding 7% in AUC, while surpassing clinical baseline models such as the UKPDS Risk Engine and BLR. These improvements provided evidence regarding the potential of GNNs to capture complex, non-linear relationships between patient attributes which may be less easily modeled by traditional statistical models and machine learning methods unable to account for relationships between patients. When compared directly with the XGBoost classifier that was recently applied to the same prediction task [15], the proposed GNN consistently delivered superior results across all metrics. While XGBoost achieved competitive AUC values, its precision, recall, and F1-score were substantially lower, indicating its limited ability to accurately identify high-risk patients under class imbalance.

Overall, the obtained performance estimates highlighted the proposed model's superiority in minimizing the false negative rate, which can be attributed to core properties of the GNN architecture that distinguish it from conventional feature-based classifiers. In particular, unlike models that treat each patient as an independent instance, the GNN explicitly models inter-patient similarity, allowing information to propagate across clinically related profiles and strengthening minority-class discrimination. In addition, the GNN learns context-aware representations: each node embedding reflects both the patient's own features and those of neighboring patients, enabling the detection of subtle, higher-order risk patterns that are not apparent from isolated feature vectors. When combined with the anomaly - detection objective - which enforces a compact representation of the majority class and treats CVD cases as structural deviations - this relational and context-aware formulation provides a principled advantage over traditional feature-based classifiers under severe class imbalance.

The integration of RuleFit as a global surrogate model enabled the generation of transparent insights into the GNN's decision-making process. The selection of RuleFit was motivated by its ability to generate explicit, human-readable rules that capture non-linear relationships and higher-order feature interactions. In contrast, feature-attribution methods such as SHAP [42] provide per-instance additive explanations based on Shapley values. Although SHAP can quantify interactions through its extended formulations, it does not present these interactions as explicit conditional logic, and its outputs typically require aggregation and post-processing to yield global patterns. RuleFit, by directly expressing decision logic in the form of IF–THEN rules, offers a more intuitive structure for clinical interpretation, reflecting the threshold-based and conditional reasoning common in guideline-driven practice. This ability to uncover multivariate, clinically meaningful patterns – including signals of residual risk in treated patients – was a key reason for selecting a rule-based surrogate over purely feature-attribution approaches.

By translating graph-derived embeddings into explicit decision rules, RuleFit facilitated understanding of how the original model combined diverse patient attributes into a unified risk assessment. The surrogate model achieved high fidelity (R-squared: 92.512%) in reproducing the predictions of the original GNN model, thereby indicating that the extracted rule sets closely approximated the underlying model behavior.

The examination of rules derived from the surrogate model across different prediction cases provided evidence regarding the model's capacity to capture clinically meaningful patterns consistent with expert knowledge. Specifically, the identified rules in the TP and TN cases linked traditional cardiovascular risk factors, such as dyslipidemia, age, and diabetes duration, with higher risk scores, while underscoring protective attributes, including adequate HDL, absence of lipid-lowering therapy, and shorter diabetes duration, as indicators of a lower risk phenotype. Beyond confirming known associations, the extracted rules in these cases also highlighted multivariate patterns that may be underappreciated in routine clinical risk scoring. For example, the combination of lipid-lowering therapy with HDL levels lower than or equal to 44.5 mg/dL was highlighted among the most influential rules for the TP prediction, indicating that the GNN may be sensitive to intermediate-risk phenotypes. In clinical terms, HDL values below 40 mg/dL in men or 50 mg/dL in women are typically considered high-risk CVD thresholds. While the identified threshold of 44.5 mg/dL does not meet these high-risk cutoffs for all patients, its occurrence in someone already on lipid-lowering therapy may indicate residual dyslipidemia and incomplete lipid control, both of which are associated with elevated long-term risk. Identifying such cases could prompt earlier intervention in T2DM, where multiple modest deviations from normal values may cumulatively increase vascular risk.

On the other hand, the analysis of rule-based explanations for the cases of misclassification revealed instances where the GNN's decision pathway diverged from clinical expectations, either due to the strong weighting of individual high-risk factors without sufficient counterbalancing from protective factors or due to the model's sensitivity to protective factors that masked underlying risks. These results may, in part, reflect constraints in the dataset, such as limited sample size or unmeasured risk factors not represented in the feature space, though model-specific biases or interaction effects could also have contributed to the observed reduced predictive accuracy in these cases. Overall, the obtained surrogate explanations largely aligned with established clinical expectations and were cross-checked with domain experts, who confirmed the clinical plausibility of the dominant patterns. They also revealed instances where the model's reasoning focused on factors that are not typically emphasized in routine clinical assessment, or diverged from conventional medical judgment. These deviations may offer complementary insights alongside conventional risk stratification while highlighting data- and model-related aspects where targeted adjustments could further enhance model performance.

From a translational perspective, the integration of these interpretable outputs into a clinical decision support system is crucial for real-world implementation. In a practical workflow, a clinician could enter a patient's data and receive two outputs: (1) an estimated CVD risk probability and (2) a concise set of rules that explain the primary factors driving that prediction. Unlike feature attribution methods including SHAP, which provide numerical importance scores requiring an additional interpretation step, the rule-based explanations are expressed in natural conditional logic (e.g., 'IF feature X is in range Y AND feature Z is true…'). This format aligns closely with clinical reasoning and can be understood with minimal additional training. Such outputs offer immediately actionable insights – for example, prompting a review of medication regimens or a focused discussion on lifestyle modification – thereby enhancing usability and reducing cognitive burden in routine practice. In this way, the framework transitions from serving solely as a risk calculator to functioning as a decision-support partner that facilitates evidence-based, individualized patient care. As a next step towards clinical translation, these components will require integration testing within EHR environments and prospective evaluation with clinicians to ensure workflow compatibility and practical utility.

Some limitations should be acknowledged. First, the model was developed and validated using data from a relatively small cohort (N=560) from a single center, which may limit generalizability and increase susceptibility to variability or distribution shifts. A nested stratified cross-validation scheme was employed to obtain an unbiased estimate of generalization performance in the absence of compatible external datasets in terms of feature availability and variable and endpoint definitions. Also, by recomputing training-derived quantities independently within each fold, this procedure prevented distribution leakage and limited fold-to-fold discrepancies arising from partitioning. Nonetheless, the observed standard deviations reflect the expected inter-fold variability inherent to a dataset of this size. Furthermore, the computational complexity of the proposed GNN framework is a practical consideration for clinical integration. While inference on a pre-trained model is fast, the reliance on a transductive learning framework necessitates full graph reconstruction and model retraining to generate predictions for new patients. A critical next step is the exploration of inductive GNN architectures that enable predictions for unseen nodes without requiring a fixed graph structure or full retraining. Finally, a formal qualitative evaluation of the rule-based explanations by an external panel of clinicians is considered essential for determining their practical utility, trustworthiness, and potential integration into clinical workflows.

6. Conclusions

This study introduced an interpretable GNN-based framework for CVD risk prediction in patients with T2DM, demonstrating both superior predictive performance and clinical transparency. The model outperformed state-of-the-art methods, achieving an AUC of 0.786, a more than 7% improvement over previous best results, and generated well-calibrated probabilities that reliably reflect observed risks. Importantly, the integration of a RuleFit surrogate provided human-readable rules that confirmed known cardiovascular risk factors while highlighting under-recognized interaction patterns, such as residual dyslipidemia under therapy.

These findings underline the potential of GNN-based approaches to enhance clinical risk stratification by capturing latent inter-patient relationships beyond traditional statistical models. While the single-center dataset size remains a limitation, the results encourage validation on larger, multi-center cohorts and exploration of inductive GNNs for scalable clinical deployment. Ultimately, the proposed framework represents a step towards integrating interpretable AI into routine diabetes care, supporting earlier, tailored interventions for CVD prevention.

CRediT authorship contribution statement

Ioannis Siachos: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria Athanasiou:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Konstantia Zarkogianni:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Anastasia C. Thanopoulou:** Resources, Data curation. **Konstantina S. Nikita:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Code availability

The source code for the experimental process is publicly available on GitHub. The repository can be accessed at https://github.com/giansiaxubi/gcn_cardiovascular_t2dm.

Declaration of competing interest

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] International Diabetes Federation, IDF diabetes atlas, 11th edn, International Diabetes Federation, Brussels, Belgium, 2025.

[2] S.J. Haffner, H. Cassells, Hyperglycemia as a cardiovascular risk factor, Am. J. Med. 115 (8) (2003) 6–11.

[3] F. Cosentino, P.J. Grant, V. Aboyans, C.J. Bailey, A. Ceriello, V. Delgado, M. Federici, G. Filippatos, D.E. Grobbee, T.B. Hansen, et al., 2019 ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: The task force for diabetes, pre-diabetes, and cardiovascular diseases of the European society of cardiology (ESC) and the European association for the study of diabetes (EASD), Eur. Heart J. 41 (2) (2020) 255–323.

[4] M.A.E.M.D. Badawy, L. Naing, S. Johar, S. Ong, H.A. Rahman, D.S.N.A.P. Tengah, C.L. Chong, N.A.A. Tuah, Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review, BMC Public Health 22 (1) (2022) 1742.

[5] A. Galbete, I. Tamayo, J. Librero, M. Enguita-Germán, K. Cambra, B. Ibáñez Beroiz, et al., Cardiovascular risk in patients with type 2 diabetes: A systematic review of prediction models, Diabetes Res. Clin. Pract. (2021) 109089.

[6] H. Amadid, N.B. Johansen, A.-L. Bjerregaard, S. Brage, K. Færch, T. Lauritzen, D.R. Witte, A. Sandbæk, M.E. Jørgensen, D. Vistisen, The role of physical activity in the development of first cardiovascular disease event: a tree-structured survival analysis of the Danish ADDITION-PRO cohort, Cardiovasc. Diabetol. 17 (1) (2018) 126.

[7] K.M. Anderson, P.M. Odell, P.W. Wilson, W.B. Kannel, Cardiovascular disease risk profiles, Am. Heart J. 121 (1) (1991) 293–298.

[8] E. Cuadrado-Godia, A.D. Jamthikar, D. Gupta, N.N. Khanna, T. Araki, M. Maniruzzaman, L. Saba, A. Nicolaides, A. Sharma, T. Omerzu, et al., Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: Logistic regression approach, Comput. Biol. Med. 108 (2019) 182–195.

[9] C. Martin, M. Vanderpump, J. French, Description and validation of a Markov model of survival for individuals free of cardiovascular disease that uses Framingham risk factors, BMC Med. Inform. Decis. Mak. 4 (1) (2004) 6.

[10] R.J. Stevens, V. Kothari, A.I. Adler, I.M. Stratton, R.R. Holman, United Kingdom Prospective Diabetes Study (UKPDS) Group, The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56), Clin. Sci. 101 (6) (2001) 671–679.

[11] J.K. Kim, S. Kang, Neural network-based coronary heart disease risk prediction using feature correlation analysis, J. Heal. Eng. 2017 (1) (2017) 2780501.

[12] S.R. Alty, S.C. Millasseau, P. Chowienczyc, A. Jakobsson, Cardiovascular disease prediction using support vector machines, in: 2003 46th Midwest Symposium on Circuits and Systems, vol. 1, IEEE, 2003, pp. 376–379.

[13] K. Zarkogianni, M. Athanasiou, A.C. Thanopoulou, K.S. Nikita, Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication, IEEE J. Biomed. Health Informatics 22 (5) (2017) 1637–1647.

[14] J. Kim, J. Lee, Y. Lee, Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree, Heal. Informatics Res. 21 (3) (2015) 167–174.

[15] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A.C. Thanopoulou, K.S. Nikita, An explainable xgboost–based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE, IEEE, 2020, pp. 859–864.

[16] J. Hu, G. Hao, J. Xu, X. Wang, M. Chen, Deep learning-based coronary artery calcium score to predict coronary artery disease in type 2 diabetes mellitus, Heliyon 10 (6) (2024).

[17] Y. Qian, P. Expert, P. Panzarasa, M. Barahona, Geometric graphs from data to aid classification tasks with Graph Convolutional Networks, Patterns (N Y) 2 (4) (2021) 100237, http://dx.doi.org/10.1016/j.patter.2021.100237, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085612/.

[18] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[19] H. Yu, G. He, W. Wang, S. Qin, Y. Wang, M. Bai, K. Shu, D. Pu, A graph neural network approach for accurate prediction of pathogenicity in multi-type variants, Brief. Bioinform. 26 (2) (2025) bbaf151.

[20] Y. Li, B. Qian, X. Zhang, H. Liu, Graph neural network-based diagnosis prediction, Big Data 8 (5) (2020) 379–390.

[21] P. Mohanraj, V. Raman, S. Ramanathan, Deep learning for parkinson's disease diagnosis: A graph neural network (GNN) based classification approach with graph wavelet transform (GWT) using protein–peptide datasets, Diagnostics 14 (19) (2024) 2181.

[22] A. Bessadok, M.A. Mahjoub, I. Rekik, Graph neural networks in network neuroscience, 2021, arXiv preprint arXiv:2106.03535.

[23] H. Mohammadi, W. Karwowski, Graph neural networks in brain connectivity studies: Methods, challenges, and future directions, Brain Sci. 15 (1) (2024) 17.

[24] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, Z. Li, MGNN: A multimodal graph neural network for predicting the survival of cancer patients, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1697–1700.

[25] G. Gogoshin, A.S. Rodin, Graph neural networks in cancer and oncology research: Emerging and future trends, Cancers 15 (24) (2023) 5858.

[26] L. Li, J. Zhou, Y. Jiang, B. Huang, Propagation source identification of infectious diseases with graph convolutional networks, J. Biomed. Informatics 116 (2021) 103720, http://dx.doi.org/10.1016/j.jbi.2021.103720, URL https://www.sciencedirect.com/science/article/pii/S1532046421000496.

[27] Q. He, Y. Bao, H. Fang, Y. Lin, H. Sun, HHAN: Comprehensive infectious disease source tracing via heterogeneous hypergraph neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, (1) 2025, pp. 291–299.

[28] Y. Gu, X. Yang, L. Tian, H. Yang, J. Lv, C. Yang, J. Wang, J. Xi, G. Kong, W. Zhang, Structure-aware siamese graph neural networks for encounter-level patient similarity learning, J. Biomed. Informatics 127 (2022) 104027, http://dx.doi.org/10.1016/j.jbi.2022.104027, URL https://www.sciencedirect.com/science/article/pii/S1532046422000430.

[29] L. Yu, M. Cheng, W. Qiu, X. Xiao, W. Lin, Idse-HE: Hybrid embedding graph neural network for drug side effects prediction, J. Biomed. Informatics 131 (2022) 104098, http://dx.doi.org/10.1016/j.jbi.2022.104098, URL https://www.sciencedirect.com/science/article/pii/S1532046422001149.

[30] A. Kazi, S. Shekarforoush, S. Arvind Krishna, H. Burwinkel, G. Vivar, B. Wiestler, K. Kortüm, S.-A. Ahmadi, S. Albarqouni, N. Navab, Graph convolution based attention model for personalized disease prediction, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 122–130, http://dx.doi.org/10.1007/978-3-030-32251-9_14.

[31] S. Parisot, S.I. Ktena, E. Ferrante, M. Lee, R.G. Moreno, B. Glocker, D. Rueckert, Spectral graph convolutions for population-based disease prediction, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2017, pp. 177–185, http://dx.doi.org/10.1007/978-3-319-66179-7_21.

[32] S. Parisot, S.I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, D. Rueckert, Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease, Med. Image Anal. 48 (2018) 117–130, http://dx.doi.org/10.1016/j.media.2018.06.001.

[33] Y. Chen, Y. Bian, B. Han, J. Cheng, How interpretable are interpretable graph neural networks?, 2024, arXiv preprint arXiv:2406.07955.

[34] J. Yang, Y. Wang, K. Chen, T. Zheng, Y. Zhou, Z. Xiao, J. Cao, M. Song, S. Liu, From GNNs to trees: Multi-granular interpretability for graph neural networks, 2025, arXiv preprint arXiv:2505.00364.

[35] C. Yang, H. Cui, J. Lu, S. Wang, R. Xu, W. Ma, Y. Yu, S. Yu, X. Kan, C. Ling, et al., A review on knowledge graphs for healthcare: Resources, applications, and promises, 2023, arXiv preprint arXiv:2306.04802.

[36] R. Xue, H. Deng, F. He, M. Wang, Z. Zhang, Trustworthy GNNs with LLMs: A systematic review and taxonomy, 2025, arXiv preprint arXiv:2502.08353.

[37] J.H. Friedman, B.E. Popescu, Predictive learning via rule ensembles, Ann. Appl. Stat. 2 (3) (2008) 916–954, URL https://www.jstor.org/stable/30245114. Publisher: Institute of Mathematical Statistics.

[38] J. Kong, W. Kowalczyk, S. Menzel, T. Bäck, Improving imbalanced classification by anomaly detection, in: Parallel Problem Solving from Nature – PPSN XVI, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 512–523, http://dx.doi.org/10.1007/978-3-030-58112-1_35.

[39] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Algorithm-level approaches, in: Learning from Imbalanced Data Sets, Springer International Publishing, Cham, 2018, pp. 123–146, http://dx.doi.org/10.1007/978-3-319-98074-4_6.

[40] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, 2019, arXiv:1901.03407 [Cs, Stat] URL http://arxiv.org/abs/1901.03407.

[41] A. Kumagai, T. Iwata, Y. Fujiwara, Semi-supervised anomaly detection on attributed graphs, in: 2021 International Joint Conference on Neural Networks, IJCNN, 2021, pp. 1–8.

[42] C. Molnar, Interpretable machine learning, 2020, Lulu. com.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958, URL http://jmlr.org/papers/v15/srivastava14a.html.

[44] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), 2016, arXiv E-Prints. arXiv:1606.08415 _eprint: 1606.08415.

[45] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2015, CoRR, arXiv:1412.6980.
[46] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.
[47] M. Ojala, G.C. Garriga, Permutation tests for studying classifier performance., J. Mach. Learn. Res. 11 (6) (2010).
[48] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, M.W. Kattan, Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology (Cambridge, Mass.) 21 (1) (2010) 128.