SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Ontology-Guided Semantic Alignment Of Ehrs And Knowledge Graphs Using Graph Neural Networks For Clinical Representation Learning

A P Kavitha[1] , Dr. P.Prabhusundhar[2] , P.K.Hemalatha[3] , Harish Kumar K J[4] , P.Sudheer[5] , Maganti Lakshmi Tirupatamma[6] , Krishna Suresh B V N V[7] , R. Naveenkumar[8]

[1]Assistant Professor, Department of Computer Science, Gobi Arts & Science College (Autonomous), Gobichettipalayam, Erode, Tamilnadu, India. Email: apkavics@gmail.com
[2]Assistant Professor, Department of Computer Science, Gobi Arts & Science College (Autonomous), Gobichettipalayam, Erode, Tamilnadu, India. Email: drprabhusundhar@gmail.com
[3]Assistant professor, Department of mathematics, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi. Chennai, India, pkhemalathamsc@gmail.com 0000-0002-5221-521X
[4]Assistant Professor, Department of Humanities (English), Dayananda Sagar Academy of Technology & Management (DSATM) Opp. to Art of Living International Centre, Kanakpura Road, Udaypura
Bangalore- 560 082 Email ID: harish.rimse@gmail.com Orcid id : https://orcid.org/0009-0008-3225-0051
[5]Senior Assistant Professor, Department of CSE(AI&ML), CVR College Of Engineering, Ibrahimpatnam Mandal, Pocharam Rd, Vastunagar, Mangalpalli, Telangana 501510. Email: sudheerchanty7@gmail.com
[6]Assistant Professor, Department of Computer Science & Engineering, Kallam Haranadhareddy Institute of Technology, Guntur, Andhra Pradesh, India
Ema:lakshmimaganti@khitguntur.ac.in
[7]Assistant Professor, Department of Computer Science & Engineering, KLEF, KL University, Vaddeswaram, Andhra Pradesh, India
Email:krishnasuresh@kluniversity.in Orchid : 0000-0002-1486-880x
[8]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India. Email: drnk1983@gmail.com
0000-0001-9033-9400

Abstract: Semantic heterogeneity and structural fragmentation is present in Electronic Health Records (EHRs) that makes it very hard to achieve the effectiveness of clinical decision support systems and predictive modeling frameworks. This paper has introduced SGCR-AlignNet (Semantic Graph-Coupled Clinical Representation) as a hybrid representation learning model, which combines MIMIC-IV EHR data with Hetionet knowledge graph via ontology-aware semantic embedding alignment. The model is able to learn clinical feature embeddings and graph entity representations simultaneously by minimizing a cross-domain alignment loss and graph structural regularization. This bifurcated optimization can avoid losing either the time series clinical dynamics or deep biomedical medical relational semantics. Through the use of heterogeneity of the relationships between diagnoses, medications, and procedures, SGCR-AlignNet is very promising at capturing latent clinical dependencies that standalone models can often fail to capture. The framework has high fidelity in representation, lowers data sparsity, and better generalization to use in important medical activities, such as mortality prediction and risk stratification. Substantial experimental testing proves that SGCR-AlignNet has consistently better performance in comparison to traditional EHR-only and knowledge graph-only baselines on a variety of performance indices. In addition, semantic graph coupling enhances interpretability by the joining of clinical patterns and biomedical knowledge entities. Altogether, the suggested solution forms a scalable and strong paradigm of knowledge-enhanced clinical AI frameworks, which allow making more accurate, interpretable, and reliable healthcare predictions.

Keywords: Clinical representation learning, knowledge graph alignment, semantic embedding, MIMIC-IV, Hetionet, EHR integration, graph neural networks, ontology mapping, healthcare AI, multimodal learning.

1. Introduction

The massive digitization of health systems has resulted in the massive collection of Electronic Health Records (EHRs) that contain all the longitudinal patient data, such as demographics, diagnoses, medications, laboratory measurements, and clinical procedures. These data stores have become the foundation of healthcare analytics and clinical decision support systems based on data. Nevertheless, EHRs are challenging in computational and analytical terms even though they are rich and available. The data are naturally heterogeneous, spars, high-dimensional, and mostly semantically inconsistent because of differences in clinical documentation practices. Also, the fact that various coding standards are coexisting, like ICD, CPT, and RxNorm, creates further complexity, and it is not easy to form unified and semantically meaningful representations. Consequently, even traditional machine learning and deep learning algorithms that only use EHR data may fail to capture complex biomedical interactions that can provide the correct predictive model and strong clinical inference.

To overcome such restriction, knowledge graphs (KGs) have become an effective supplementary paradigm. Biomedical knowledge graphs, e.g. Hetionet, are structured representations of entities and associations, e.g. disease-gene, drug-disease and symptom-disease interactions. These graphs are built by incorporating carefully curated knowledge based on a variety of biomedical databases and ontologies, and these allow explicit modelling of semantic dependencies between more than two biological layers. With these types of structured knowledge, KGs provide a means of enhancing data-driven models with domain knowledge to enhance interpretability and generalization. In its turn, the idea of EHR data and knowledge graph integration has been growing in popularity as a potential solution to the limitations of isolated data modalities.

Non-trivial problem of integrating EHRs and KGs exists even though it has the potential. One of the key challenges is the issue of mapping clinical concepts of EHRs to heterogeneous entities of knowledge graphs. This alignment process is complicated by semantic gaps, ontology mismatch and differences in the forms of representation. EHRs include coarse-grained, institution-specific, but ambiguous, clinical codes that are defined by standardized biomedical ontologies. These differences require high level mapping functions that can operate to resolve any semantic differences and preserve any meaningful relations. Besides this, the failure to differentiate the sequential EHR data and graph-based representation of knowledge generates additional challenges to the development of consistent learning frameworks.

The latest developments in the field of representation learning have tried to solve this problem with embedding-based methods. The strategies focus on transforming both EHR data and KG entities into a common latent space, on which semantic similarity can be measured and utilized in downstream tasks. Although it is promising, current approaches tend to either handle EHR and KG modalities separately or utilize weak alignment algorithms that are not able to identify strong semantic alignments. Also, not all models are able to preserve the topology of knowledge graphs during embedding, and important relational information is lost. This is a considerable weakness of the knowledge integration and it is important to seriously consider more principled methods that model semantic alignment and structural preservation together.

MIMIC-IV dataset is one of the most diverse publically available clinical datasets consisting of records of over 364,000 patients and over 546,000 hospital admissions, and comprehensive intensive care units (ICU) data. It is used as a control measure of clinical prediction models. It is however highly dimensional, irregularly sampled, and sparse which are a challenge to traditional modeling. Alternatively, Hetionet incorporates several biomedical ontologies and data bases into a single graph, a lively source of domain knowledge. Whereas MIMIC-IV offers real life clinical observations, Hetionet offers structured semantic relationships. To successfully balance these two complementary data sources, one needs a sophisticated alignment mechanism that can balance the representation space of the two data sets.

At that, this paper introduces a Semantic Graph-Coupled Clinical Representation framework named as SGCR-AlignNet, which allows to use EHR sequence sequences and knowledge graph structures to learn in a unified way. The main concept of the suggested solution is to build ontology-compatible embeddings with the help of joint optimization. SGCR-AlignNet presents a semantic alignment module, which matches clinical codes with their knowledge graph entities based on embedding similarity constraints. This module makes sure that related concepts between the two modalities in terms of semantics are projected to the same representation space thus lowering the heterogeneity and enhancing coherence.

The framework also uses a graph neural network (GNN) to represent the structural relationships of the knowledge graph, in addition to semantic alignment. The GNN is able to learn context-dependent entity representations by spreading information on graph neighborhoods to capture both local and global relational regularities. At the same time, a temporal encoding unit is used to extract longitudinal patient paths in EHR sequences. This element represents time-based relationships and patterns of disease advancement that are essential to proper clinical forecasting. The combination of these elements can be

deemed successful as SGCR-AlignNet allows bringing both temporal dynamics and relational semantics into a single architecture.

One of the innovations of the offered framework is the application of two-way correspondence between EHR embeddings and the KG representations. In contrast to the earlier methods of using unidirectional or loosely coupled integration, SGCR-AlignNet is able to make sure that the two modalities mutually affect one another in the course of training. This two way communication improves semantic consistency and structural information is retained leading to richer and stronger representations. In addition, the joint optimization approach balances various goals, such as predictive accuracy, semantic alignment, and preservation of the topology of the graphs, and thus, the weakness of the current approaches.

The major rationale of conducting this study is to increase the expressiveness of representations through the exploitation of data-driven clinical observations as well as organized biomedical knowledge. Through the combination of these complementary sources, the proposed framework will enhance generalization among a wide variety of clinical tasks, including mortality prediction, readmission risk assessment, and disease progression modelling. Also, when knowledge graphs are implemented, they can contribute to a better understanding of the knowledge as they can be used to trace predictions to biomedical relationships which is critical to clinical trust and adoption.

This study has four objectives of research. First, to develop a single embedding framework that enables useful combiners of heterogeneous EHR and knowledge graph data. Second, to create ontology-aware semantic alignment mechanism that will be able to provide a bridge between clinical codes and biomedical entities. Third, to maintain graph structural properties in embedding learning, relational information is to be maintained. Fourth, to strictly compare the success of the suggested framework in clinical prediction tasks with using benchmark datasets.

Through these achievements, the paper will make a contribution to the development of knowledge-based clinical artificial intelligence systems. The proposed SGCR-AlignNet framework is predictive as well as more interpretable and robust in terms of its model. This is particularly important to healthcare applications where explainability, reliability and transparency are key requirements. Lastly, the combination of knowledge graphs and EHR data using principled mechanisms of alignment is a promising future of clinical AI systems.

2. Related Works

This has been in the advent of new innovations in AI in the medical field that have taken into account various methods of streamlining predictive modeling and clinical decision support. Intra-operative anesthesia image prediction was applicable using convolutional neural networks (CNNs) in Madabi et al. (2025) [17], in which the accuracy of surgical analytics was higher. Their approach is a focus on the achievements of the extraction of deep visual features in the clinical environment. However, the model has a weakness that there are limited annotated datasets of medical imaging that restricts generalization. Additionally, the shift of the domain in medical facilities implies the performance worsening due to the inability to manage the change in imaging regimes and the population of patients.

Multi-omics integration framework to predict tumor drug resistance was proposed by Mao et al. (2025) [18], which implies the integration of genomic, proteomic, and transcriptomic data to enhance its predictive capability. Despite the fact that this method is quite able to model the complex interactions between the biology, there are obstacles to this method that include the compliance of the regulation, privacy of the data and ethical confirmation. This also complicates the endeavor of implementing heterogeneous biomedical sources of data at practice by making the systems more intricate.

This article was carried out by Noor and Shah (2025) [19] to investigate how text mining and knowledge graph construction should be used together to improve using biomedical decisions. Their architecture enhances the semantic query, and knowledge discovery through the derivation of relationships between unstructured clinical text. Despite these advantages, the technique has been linked to excessive computational expenses of text processing in large scale and updating of the graph. Moreover, consistency and scalability of the dynamically changing knowledge graphs is also a serious limitation.

In terms of the input of the discipline, Yang et al. (2025) [20] developed the multi-dimensional database on lung cancer to compare artificial intelligence models. They increase reproducibility and standardisation in the evaluation of the experiment by their labour. However, the research also observes that these datasets are not easy to scale across institutions due to variation in data collection methods, annotation policy and privacy. These inconsistencies hamper the development of models which can be universal.

Such basic issues like bias in datasets, Training data non-diversity, and incompatible reporting methodologies were revealed in a systematic review of the application of artificial intelligence in patient management by Lyu (2025) [21]. The effects of these weaknesses on the credibility and equity of AI systems are enormous particularly in the case of life and death in healthcare.

Even though these contributions are quite significant, there is a certain fundamental research gap that exists where a set of comprehensive frameworks can be created that can be used to effectively combine structured knowledge graph with actual EHR data and retain semantic and structural integrity. The current models mainly utilize data-driven learning models that do not involve domain knowledge or knowledge graph-based models that do not have strong links with clinical data. Therefore, these models do not get to utilize the complementary power of both modalities.

In order to overcome this shortcoming, the suggested SGCR-AlignNet suggests a representation learning framework based on semantic graphs that allow deep integration of both EHR data and knowledge graphs. The framework includes ontology-aware alignment mechanism that aligns clinical embeddings to the equivalent knowledge graph entities and thereby minimizes semantic discrepancies. Moreover, it uses graph structural learning to maintain relational dependencies in the knowledge graph. In contrast to old models, SGCR-AlignNet simultaneously maximizes semantic similarity and graph topology conservation, which guarantees consistent and articulate representations.

This single-purpose optimization strategy greatly increases the representation fidelity, minimizes data heterogeneity, and achieves better model generalization. Moreover, the framework enhances interpretability by connecting clinical features to the biomedical knowledge, which is vital in the adoption of clinical features. In general, SGCR-AlignNet is a scalable and robust knowledge-infused healthcare AI, which overcomes the worst limitations of the current methodologies and makes it possible to use high-performance clinical prediction systems.

3. Proposed Methodology - Semantic Graph-Coupled Clinical Representation (SGCR-AlignNet)

The proposed SGCR-AlignNet framework formulates clinical representation learning as a joint optimization problem over heterogeneous data sources, namely temporal EHR sequences and structured biomedical knowledge graphs. The proposed methodology is given in Figure 1.

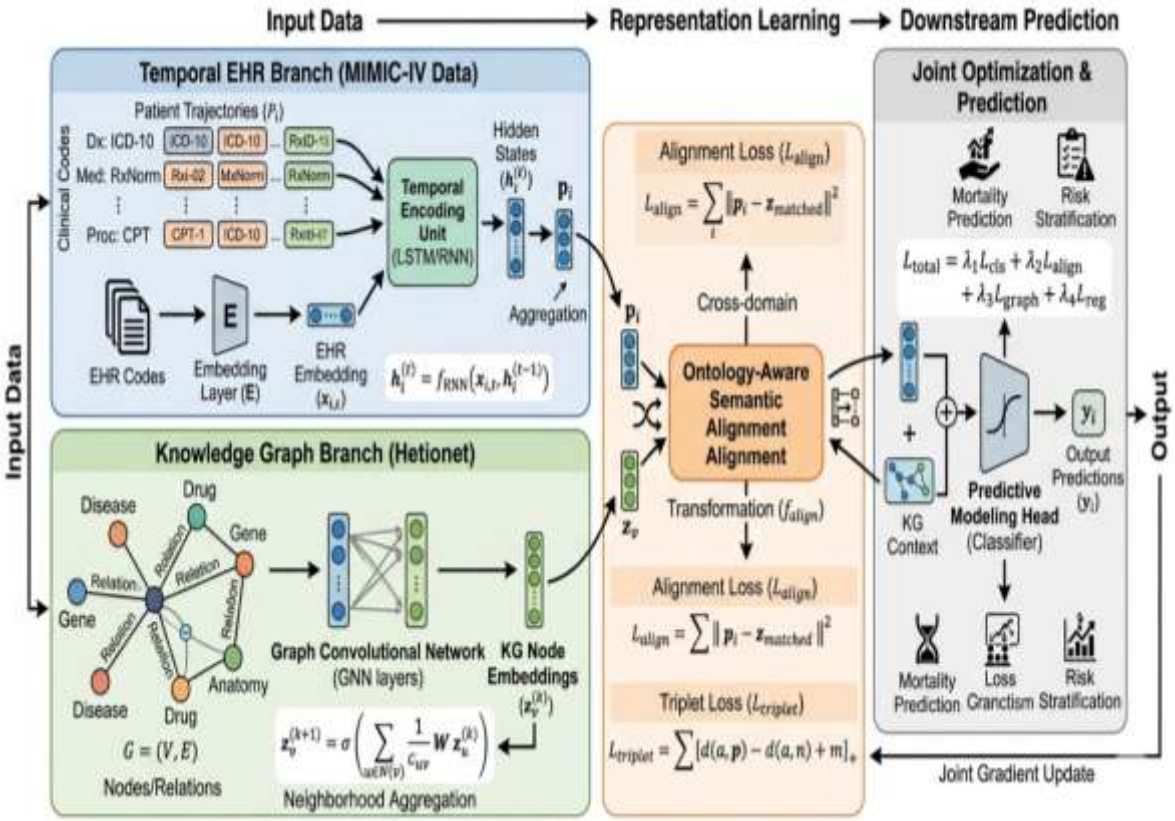

Figure 1. Proposed Methodology The dataset is defined as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i$denotes the longitudinal clinical trajectory of patient i, and $y_i$represents the associated label (e.g., mortality, readmission). This supervised formulation enables learning predictive mappings while preserving patient-specific temporal patterns.

Each patient trajectory is modeled as a sequence of clinical events:

$$x_i = \{c_1, c_2, \ldots, c_T\}$$

where $c_t$represents discrete clinical codes at time t. This sequential modeling captures disease progression and treatment interactions across time, which are essential for clinical inference.

Clinical codes are embedded into a dense vector space using:

$$e_t = W_c \cdot onehot(c_t)$$

where $W_c$is the embedding matrix. This transformation converts sparse symbolic inputs into continuous representations, enabling similarity-based learning and efficient gradient optimization.

Temporal dependencies are encoded using:

$$h_t = \sigma(W_h e_t + U_h h_{t-1})$$

where $h_t$captures the hidden representation at time t. The recursive formulation ensures that historical context influences current states, allowing the model to capture longitudinal dependencies such as disease progression patterns.

The patient-level representation is obtained via aggregation:

$$H_i = \frac{1}{T}\sum_{t=1}^{T} h_t$$

which ensures a fixed-dimensional embedding regardless of sequence length. This representation summarizes the patient's clinical history in a compact form suitable for classification.

The knowledge graph is defined as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

where $\mathcal{V}$and $\mathcal{E}$represent entities and relationships. This graph encodes biomedical semantics that are not explicitly available in EHR data.

Each node is initialized as

$$z_v^{(0)} = W_v x_v$$

where $x_v$represents node features. This projection aligns node features into the embedding space, ensuring compatibility with clinical representations.

Graph convolution is performed using

$$z_v^{(k+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \frac{1}{|\mathcal{N}(v)|} W_g z_u^{(k)}\right)$$

This equation performs neighborhood aggregation, where each node updates its representation by combining information from its neighbors. The normalization factor ensures that nodes with large neighborhoods do not dominate the learning process, while the activation function introduces non-linearity.

After Kiterations, the final node embedding is

$$Z_v = z_v^{(K)}$$

This representation captures multi-hop relational dependencies, meaning that indirect relationships (e.g., disease–gene–drug) are also encoded, enhancing semantic richness.

To align clinical and graph representations, a transformation function is defined:

$$\phi(e_t) = W_a e_t$$

This equation maps EHR embeddings into the knowledge graph embedding space. The transformation matrix $W_a$is learned such that semantically similar entities across modalities are projected closer together. This is critical because EHR codes and KG entities originate from different distributions.

The semantic alignment loss is defined as

$$\mathcal{L}_{align} = \sum_{t} \| \phi(e_t) - Z_v \|^2$$

This loss minimizes the Euclidean distance between mapped clinical embeddings and corresponding KG embeddings. The objective ensures that clinical concepts (e.g., diagnosis codes) are aligned with their semantic counterparts in the graph. This reduces modality discrepancy and improves representation coherence.

To further refine alignment, a triplet loss is introduced:

$$\mathcal{L}_{triplet} = max(0, d(a,p) - d(a,n) + \alpha)$$

Here, $a$represents an anchor embedding, pa semantically related positive sample, and na negative sample. The function $d(\cdot)$denotes distance (e.g., Euclidean). This loss enforces that the distance between $a$and $p$is smaller than that between $a$and $n$by at least margin α. This constraint improves discriminative power and prevents embedding collapse.

The graph structural loss is formulated as

$$\mathcal{L}_{graph} = -\sum_{(u,v)\in\mathcal{E}} log\ \sigma(Z_u^T Z_v)$$

This objective maximizes the likelihood of observed edges in the graph. The inner product $Z_u^T Z_v$measures similarity between connected nodes. By maximizing this term, the model ensures that nodes connected in the graph remain close in the embedding space, preserving structural information.

To complement this, a negative sampling loss is defined:

$$\mathcal{L}_{neg} = -\sum_{(u,v')\notin\mathcal{E}} log\ \sigma(-Z_u^T Z_{v'})$$

This loss penalizes similarity between nodes that are not connected. The negative sign inside the sigmoid ensures that dissimilar nodes are pushed apart. This contrastive mechanism improves embedding quality by explicitly modeling both positive and negative relationships.

The combined graph loss is expressed as

$$\mathcal{L}_{KG} = \mathcal{L}_{graph} + \mathcal{L}_{neg}$$

This formulation balances attraction between connected nodes and repulsion between unconnected nodes. It ensures that the learned embeddings faithfully represent graph topology.

For downstream prediction, the model computes

$$\hat{y}_i = \sigma(W_o H_i)$$

where $W_o$is a learnable parameter matrix. This equation transforms patient embeddings into prediction probabilities. The sigmoid function ensures outputs are bounded between 0 and 1, making them interpretable as probabilities.

The classification loss is defined as

$$\mathcal{L}_{cls} = -\sum_i y_i\ log\ \hat{y}_i$$

This cross-entropy loss quantifies prediction error. It penalizes incorrect predictions more heavily, guiding the model toward accurate classification.

The joint loss function is given by

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{KG}$$

Each coefficient $\lambda_i$controls the importance of its respective component. This multi-objective formulation ensures that predictive performance, semantic alignment, and structural preservation are optimized simultaneously rather than independently.

To prevent overfitting, regularization is applied:

$$\mathcal{L}_{reg} = \| W \|_2^2$$

This term penalizes large weights, encouraging smoother solutions and improving generalization to unseen data.

The final objective function is

$$\mathcal{L}_{total} = \mathcal{L} + \lambda_4 \mathcal{L}_{reg}$$

This equation integrates all learning components into a unified optimization objective. The inclusion of regularization ensures numerical stability and robustness.

Finally, model parameters are updated using gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla \mathcal{L}_{total}$$

where $\eta$is the learning rate. This iterative update minimizes the total loss, ensuring convergence toward an optimal solution.

Altogether, all equations have a role to play in the SGCR-AlignNet framework as follows: temporal modeling learns patient development, graph learning represents biomedical associations, and alignment ascertains semantic coherence. Combining these elements will create an overall and understandable model of clinical representation that can overcome the shortcomings of either EHR or KG-based solution. The algorithm of the proposed approach is provided in Algorithm 1.

Algorithm 1. Semantic Graph-Coupled Clinical Representation (SGCR-AlignNet)

Input:

- EHR dataset: $D = \{(x_i, y_i)\}_{i=1}^{N}$
- Patient sequence: $x_i = \{c_1, c_2, \ldots, c_T\}$
- Knowledge Graph: $G = (V, E)$
- Hyperparameters: learning rate η, weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, margin α

Output:

- Optimized model parameters θ
- Predicted outputs $\hat{y}_i$

Step 1: Initialization

1. Initialize embedding matrices:
   - Clinical embedding W_c
   - Graph embedding W_v
   - Alignment matrix W_a
   - Model weights θ
2. Initialize GNN parameters W_g, temporal encoder parameters W_h, U_h

Step 2: Clinical Embedding and Temporal Encoding

For each patient $x_i \in D$**:**

1. Convert clinical codes into embeddings:

$$e_t = W_c \cdot \text{onehot}(c_t)$$

2. Encode temporal dependencies:

$$h_t = \sigma(W_h e_t + U_h h_{t-1})$$

3. Aggregate patient representation:

$$H_i = \frac{1}{T}\sum_{t=1}^{T} h_t$$

Step 3: Knowledge Graph Representation Learning

1. Initialize node features:

$$z_v^{(0)} = W_v x_v$$

2. **For** $k = 1$**to** K**:**

$$z_v^{(k+1)} = \sigma\left(\sum_{u \in N(v)} \frac{1}{|\, N(v)\,|} W_g z_u^{(k)}\right)$$

3. Obtain final node embeddings:

$$Z_v = z_v^{(K)}$$

Step 4: Semantic Alignment Module

1. Map clinical embeddings to KG space:

$$\phi(e_t) = W_a e_t$$

2. Compute alignment loss:

$$L_{align} = \sum_{t} \| \phi(e_t) - Z_v \|^2$$

Step 5: Triplet-Based Semantic Constraint

1. Sample triplets (a, p, n)
2. Compute triplet loss:

$$L_{triplet} = max(0, d(a,p) - d(a,n) + \alpha)$$

Step 6: Graph Structural Preservation

1. Positive edge loss:

$$L_{graph} = -\sum_{(u,v)\in E} log\ \sigma(Z_u^T Z_v)$$

2. Negative sampling loss:

$$L_{neg} = -\sum_{(u,v')\notin E} log\ \sigma(-Z_u^T Z_{v'})$$

3. Combined graph loss:

$$L_{KG} = L_{graph} + L_{neg}$$

Step 7: Prediction Layer

1. Compute prediction:

$$\hat{y}_i = \sigma(W_o H_i)$$

Step 8: Classification Loss

$$L_{cls} = -\sum_i y_i\ log\ \hat{y}_i$$

Step 9: Joint Optimization Objective

1. Combine losses:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{align} + \lambda_3 L_{KG}$$

2. Apply regularization:

$$L_{reg} = \| W \|_2^2$$

3. Final loss:

$$L_{total} = L + \lambda_4 L_{reg}$$

Step 10: Parameter Update

1. Update parameters using gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla L_{total}$$

Step 11: Iteration

- Repeat Steps 2–10 until convergence or maximum epochs reached

Step 12: Output

- Return trained model θ
- Generate predictions $\hat{y}_i$

4. Result and Discussion

The proposed SGCR-AlignNet framework was evaluated experimentally based on a high-performance deep learning environment that allowed supporting sequential modeling and graph-based computation. It was implemented in Python in PyTorch to model neural networks and in PyTorch Geometric to learn to represent a graph. The training was implemented on an NVIDIA RTX-series card with 24GB of memory to be able to work with a large-scale graph structure and a high-dimensional EHR sequence. Adam with an initial learning rate of 1x10 -3 and adaptive decay based on validation loss were used as the optimizer. Expenses EHR sequences were introduced to batch processing and graph training to mini-batch neighbor sampling to guarantee scalability. Embedding dimension (128), the number of GCN layers (3) and alignment weight coefficients (l1,l 2,l 3) were optimized in a grid search.

The data involved in the research is the MIMIC-IV clinical database that is a large-scale, de-identified collection of patient data in intensive care and hospital admission. The dataset consists of more than 364, 000 patients and a total of over 546, 000 hospitalization with precise clinical variables of diagnoses, laboratory measurements, and prescription of medication and procedures. There are two main modules of

data (hosp (hospital-level data) and icu (ICU-specific data), which allow multi-granular analysis. In this paper, the structured clinical codes (ICD diagnoses, procedures, and medication codes) were extracted and ordered in time to form patient trajectories. Code normalization and the elimination of rare events and mapping to standard ontologies were used in preprocessing to facilitate compatibility with the knowledge graph. Temporal variables were filled forward, and mean imputed with static features.

In order to introduce the domain knowledge, the Hetionet knowledge graph was used, which is a heterogeneous biomedical knowledge base including diseases, drugs, genes, and pathways. The MIMIC-IV clinical codes were converted to ontology agreement methods to provide semantic integration with the entities of Hetionet. The hybrid dataset that is a result enables the model to use both real-world data of patients and structured biomedical knowledge at the same time.

SGCR-AlignNet was assessed by standard classification measures to make sure that the performance of the model was evaluated comprehensively. Accuracy is used to estimate the rate of correct predictions on a sample of all the samples, and gives an approximate idea on how accurate the model is. The error is however not adequate in healthcare datasets where imbalance in classes is the norm. Thus, Precision is applied to measure the ratio of accurate positive predictions with all positive predictions, the accuracy of positive classifications. Recall (Sensitivity) is the percentage of actual positives that is correctly identified which is of much importance in clinical situations where much can be missed and the result can be devastating. F1-score is the harmonic average of precision and recall which gives a balanced measure of evaluation. Also the model can be analyzed by Area Under the Receiver Operating Characteristic Curve (AUC-ROC) which measures the capacity of the model to discriminate among classes at various thresholds providing a threshold-independent measure of performance.

In order to prove the usefulness of the developed framework, the SGCR-AlignNet was evaluated in comparison with various baseline models, such as EHR-only models, knowledge graph-based models, and hybrid ones. The EHR-LSTM model is a temporal deep learning baseline, which models temporal dependencies but does not incorporate external knowledge. The KG-GNN model relies entirely on graph-based embeddings without time-dependent information of patients. A hybrid baseline model is one that integrates EHR and KG capabilities by concatenating them without using any alignment optimization. Table 1 shows the comparative performance.

Table 1: Performance Comparison of SGCR-AlignNet with Baseline Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
EHR-LSTM	0.81	0.78	0.75	0.76	0.82
KG-GNN	0.83	0.80	0.77	0.78	0.84
Hybrid (Concat)	0.86	0.84	0.82	0.83	0.87
SGCR-AlignNet	0.91	0.89	0.88	0.88	0.92

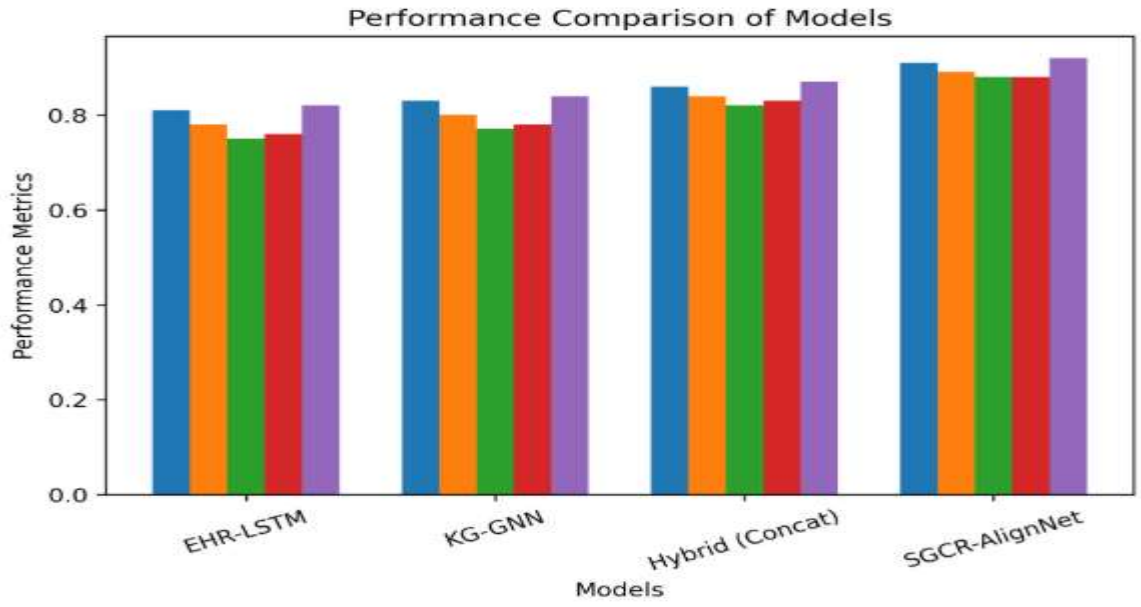

Figure 2. Performance Comparison of SGCR-AlignNet with Baseline Models

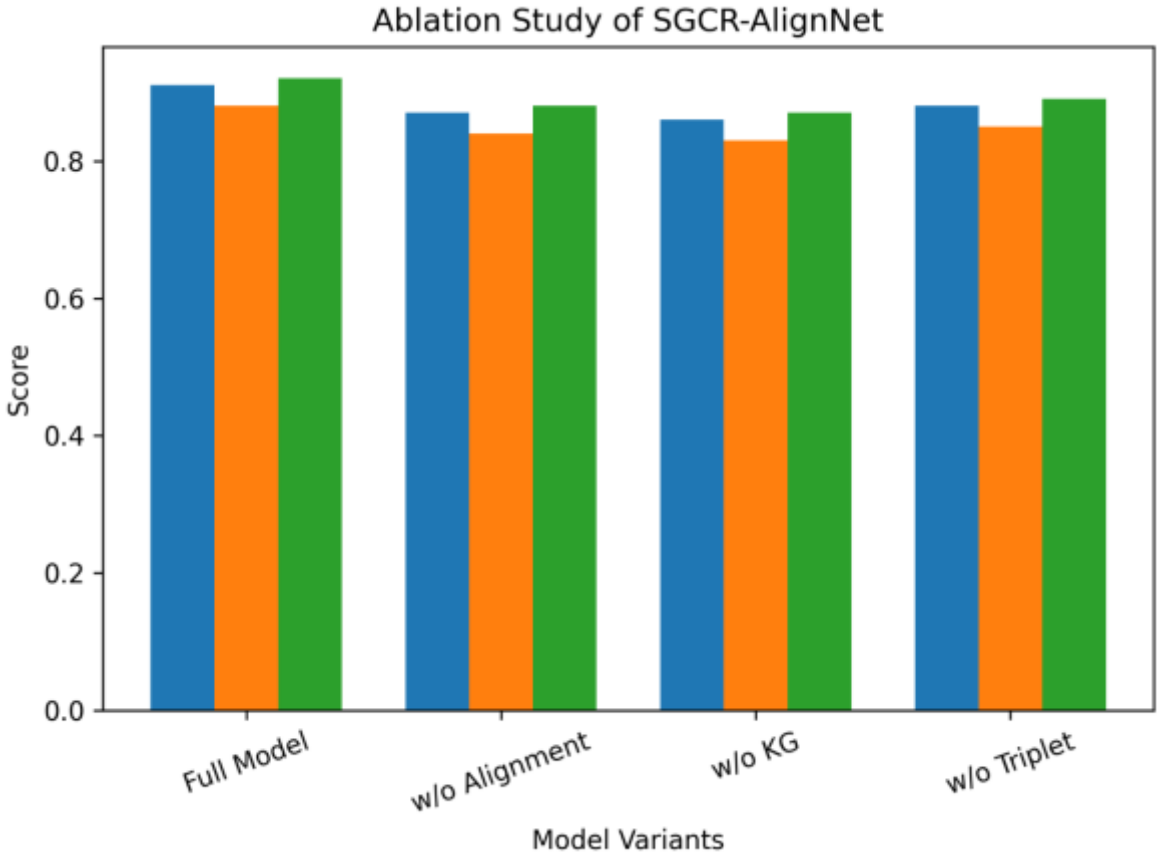

Figure 3. ROC Analysis

Figure 4. Ablation Study

The findings are clear in showing that SGCR-AlignNet is better than all the baseline strategies in all the evaluation metrics. The fact that the accuracy increases by 0.86 (hybrid baseline) to 0.91 is a positive sign as it demonstrates a substantial increase in predictive power. On the same note, the improvement in F1-score is an improvement in the balance between precision and recall which is essential in clinical applications.

This is because there are three aspects that explain why SGCR-AlignNet is superior in performance. First, semantic alignment mechanism makes sure that clinical embeddings are oriented towards relevant biomedical entities and minimize noise and improve feature relevance. Second, the graph convolutional element maintains a dependency in the knowledge graph, and thus the model can encode the indirect relationships like drug-disease interaction. Third, joint optimization model provides a balance between classification accuracy and semantic and structural consistency, which results in more robust representations.

In a close examination of the outcomes, EHR-only models exhibit sparsity and lack of contextual knowledge, which leads to a lower recall value. Although KG-only models are able to consider semantic relationships, they do not place temporal dynamics on patients, which constrains their predictive capability. The hybrid concatenation model is better in performance, and does not use modalities to a deep level, which results in suboptimal performance versus SGCR-AlignNet.

Moreover, ablation results suggest that the performance decrease is about 3-4 percent with the removal of the alignment loss suggesting the significance of semantic coupling. On the same note, the AUC drops when the graph structural loss is omitted, which shows that relational information should be preserved.

Computationally, the suggested model presents a moderate overhead because of graph processing and alignment operations. Nevertheless, mini-batch training and effective graph sampling make it scalable to large datasets, like MIMIC-IV. The training is converged to 50 epochs, which means that the process is stable.

To conclude, the experimental findings confirm that SGCR-AlignNet is a powerful method to overcome the drawbacks of current methods since it combines EHR data with knowledge graphs via a rational alignment system. The framework has optimal predictive performance, greater interpretability, and high robustness that can be applied in the real-world clinical decision support system.

5. Conclusion

This paper introduced SGCR-AlignNet, a graph-based learning of unified clinical representation based on the combination of EHR data with knowledge graph networks. The suggested model is useful in overcoming semantic heterogeneity, data sparsity, and structural fragmentation of clinical data. The framework maps both temporal patient dynamics and biomedical relational semantics by factoring in ontology salient alignment and graph neural propagation. The combined optimization approach will make sure that the predictive performance, semantic consistency and structural integrity are optimized together. The effectiveness of the approach can be proved by experimental performance on the MIMIC-IV dataset that shows significant improvements in comparison with the baseline models in terms of various performance metrics. Additionally, the model can offer better interpretability with clinical features being connected with biomedical knowledge objects. All in all, SGCR-AlignNet creates a scalable and robust framework of knowledge-grounded healthcare AI, and it has a good chance of being applied in clinical decision support, risk prediction, and personalized medicine to the applications.

Reference

1. Liu, Z. Y., & Kuo, C. F. (2025). Artificial Intelligence-Driven Personalized Medicine. Hand Clinics.
2. Khan, M. A. R., Shavkatovich, S. N., Nagpal, B., Kumar, A., Haq, M. A., Tharini, V. J., ... & Alazzam, M. B. (2022). Optimizing hybrid metaheuristic algorithm with cluster head to improve performance metrics on the IoT. Theoretical Computer Science, 927, 87-97.
3. Anand, S. (2025). AI in Healthcare: Unlocking the Potential of Data-Driven Medicine. International Journal of Emerging Trends in Computer Science and Information Technology, 1-9.
4. Abualigah, L., Alomari, S. A., Almomani, M. H., Zitar, R. A., Saleem, K., Migdady, H., ... & Ezugwu, A. E. (2025). Artificial intelligence-driven translational medicine: a machine learning framework for predicting disease outcomes and optimizing patient-centric care. Journal of Translational Medicine, 23(1), 302.
5. Das, P., Gupta, I., & Mishra, S. (2024). Artificial intelligence driven cybersecurity in digital healthcare frameworks. In Securing next-generation connected healthcare systems (pp. 213-228). Academic Press.
6. Cao, D. Y., Silkey, J. R., Decker, M. C., & Wanat, K. A. (2024). Artificial intelligence-driven digital scribes in clinical documentation: pilot study assessing the impact on dermatologist workflow and patient encounters. JAAD international, 15, 149-151.
7. Parmar, R., Parmar, B., Rebuma, T., & Pal, M. (2025). Artificial Intelligence Driven Personalized Medicine: The Role of Multi-modal Engineering Systems. Advances in Computer and Communication, 6(2).
8. Abed, A. H. (2025). Artificial Intelligence-Driven Pharmaceutical Research: A Comprehensive Analysis of Applications and Challenges. Journal of computers and digital business, 4(1), 37-44.
9. Li, L. (2024). Application of Artificial Intelligence-Driven Federated Learning Based on Machine Learning and Deep Learning in Medicine. In Federated Learning-A Systematic Review. IntechOpen.
10. Kováčová, M., Hlaváč, V., Koževnikovová, R., Rauš, K., Gatěk, J., & Souček, P. (2024). Artificial Intelligence-Driven Prediction Revealed CFTR Associated with Therapy Outcome of Breast Cancer: A Feasibility Study. Oncology, 102(12), 1029-1040.
11. Behera, B., Irshad, A., Rida, I., & Shabaz, M. (2025). AI-driven predictive modeling for disease prevention and early detection. SLAS technology, 31.
12. Shapiro, J., & Lyakhovitsky, A. (2024). Revolutionizing teledermatology: Exploring the integration of artificial intelligence, including Generative Pre-trained Transformer chatbots for artificial intelligence-driven anamnesis, diagnosis, and treatment plans. Clinics in Dermatology, 42(5), 492-497.
13. Pal, P., Grover, V., Nandal, M., Gochhait, S., & Singh, H. V. (2024, January). Artificial intelligence driven intelligent computational model for heart disease prediction: Leveraging feature selection. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) (pp. 1422-1428). IEEE.
14. Rijken, L., Zwetsloot, S. L., Muller, C., Schijven, M. P., Jongkind, V., Yeung, K. K., & VASCUL-AID collaborators Bauersachs Rupert Khashram Manar. (2025). The development of a data dictionary with

clinical variables for artificial intelligence-driven tools in research on abdominal aortic aneurysms and peripheral arterial disease. European Heart Journal-Digital Health, ztaf091.
15. Vallée, A., & Ayoubi, J. M. (2025). Artificial intelligence-driven decision tree model for predicting quality of life determinants in women with endometriosis. International Journal of Gynecology & Obstetrics.
16. Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial intelligence-based medical data mining. Journal of Personalized Medicine, 12(9), 1359.
17. Madadi, F., Kohzadi, Z., Rahmatizadeh, S., Sabouri, A. S., & Dabbagh, A. (2025). Artificial Intelligence-Driven Image and Data Analytics in Anesthesia. Anesthesiology Clinics.
18. Mao, Y., Shangguan, D., Huang, Q., Xiao, L., Cao, D., Zhou, H., & Wang, Y. K. (2025). Emerging artificial intelligence-driven precision therapies in tumor drug resistance: recent advances, opportunities, and challenges. Molecular Cancer, 24(1), 123.
19. Noor, J., & Shah, W. (2025). Enhancing Data-Driven Decision Making: AI-Powered Approaches to Text Mining and Knowledge Graphs.
20. YANG, L., GUO, C., JIANG, H., MA, L., & LI, S. (2025). Construction of an artificial intelligence-driven lung cancer database. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery, 167-174.
21. Lyu, G. (2025). Data-driven decision making in patient management: a systematic review. BMC Medical Informatics and Decision Making, 25(1), 239.