A System and Method for Concurrent Multi-Disease Prediction via Knowledge-Guided Graph Neural Networks and Ensemble Learning

[1]Harsh Sawant, [2]Bhavana Vaddadi, [3]Jessica Suthar, [4]Dev G Rathor
[1234]Department of Computer Engineering
[1234]Shree L. R. Tiwari College of Engineering Mumbai, India
[1]harsh.b.sawant@slrtce.in, [2]bhavana.v.vaddadi@slrtce.in, [3]jesica.s.suthar@slrtce.in, [4]dev.g.rathor@slrtce.in

***Abstract*—Chronic disease management reveals a fundamental gap in clinical AI: most deployed prediction systems evaluate a single disease in isolation, despite overwhelming clinical evidence that patients develop and manage multiple interrelated conditions simultaneously. This paper presents an architecture for concurrent multi-disease prediction that addresses three core limitations of prior work. First, the patient health state is modeled as a temporal sequence using an LSTM-based encoder that captures disease trajectory alongside current biomarker values. Second, a hybrid knowledge graph is constructed from SNOMED-CT and DisGeNET ontology priors overlaid with data-driven co-occurrence weights. Third, the tabular ensemble and Graph Neural Network are fused through a joint co-training loss enabling shared gradient flow. Experimental results on three benchmark datasets demonstrate strong performance: the Heart Disease model achieves F1=0.8923 and ROC-AUC=0.9310; Parkinson's attains accuracy=0.9231 with perfect recall (1.000); and the Diabetes model achieves ROC-AUC=0.8388. Comorbidity analysis further confirms a +5.17% average heart disease risk elevation in the diabetic cohort, validating inter-disease interaction modeling.**

***Index Terms*—multi-disease prediction, graph neural networks, knowledge graph, temporal encoder, LSTM, co-training, SHAP explainability, multi-label classification, chronic disease, clinical decision support.**

I. Introduction

Chronic non-communicable diseases account for the majority of global morbidity and healthcare expenditure. Their clinical challenge is not merely their prevalence, but their tendency to cluster and mutually reinforce one another. A patient with poorly controlled blood glucose is statistically far more likely to develop hypertension, chronic kidney disease (CKD), and cardiovascular complications not by coincidence, but because insulin resistance, systemic inflammation, and endothelial dysfunction drive all of them simultaneously. This biological reality is well understood by clinicians yet almost entirely ignored by deployed clinical AI tools.

The dominant paradigm in medical machine learning remains the single-disease model: select a curated dataset, define a binary target, train a classifier, and report accuracy.

The result is a fragmented ecosystem where clinicians must navigate separate applications for each condition, re-enter patient data repeatedly, and reconcile disconnected predictions that carry no awareness of each other. This is clinically dangerous when a model's confident 'low diabetes risk' output fails to account for concurrent kidney dysfunction that, in a relational system, would substantially revise that estimate.

This paper proposes that the correct unit of prediction is not a single disease but a disease graph a structured representation of how chronic conditions relate to one another, through which a patient's clinical signal propagates to produce simultaneous, mutually informed risk estimates. The system integrates three key improvements: (1) a temporal encoder treating patient health as a time-series trajectory; (2) a hybrid knowledge graph combining biomedical ontology structure with data-driven edge weights; and (3) a joint co-training loss fusing tabular and graph models during training rather than post-hoc.

The remainder of this paper is structured as follows. Section II reviews prior work. Section III formalizes the prediction problem. Section IV presents the methodology. Section V describes implementation. Section VI reports experimental results. Section VII concludes.

II. Literature Review

A. Single-Disease and Multi-Disease Prediction

Gaurav et al. [2] demonstrated that rarity-based symptom weighting substantially improves accuracy, achieving 97% on 4,500 patients using an RF-LSTM-SVM pipeline, but treated each disease independently with no inter-disease signal transfer. Venkatesh [5] and Ahmad Beg et al. [6] both deployed multi-disease Streamlit applications combining SVM, Logistic Regression, and deep learning, effectively a unified interface hiding fully independent models. Rehman et al. [8] replicated this pattern with stronger evaluation (98.75% accuracy on kidney disease using RFC) but reached the same structural conclusion: aggregating single-disease models does not constitute a multi-disease model.

B. Temporal and Deep Learning Approaches

Razavian et al. [9] provided rigorous evidence that temporal modeling benefits multi-disease prediction at scale, predicting 171 disease onsets across 298,000 patients using ensemble LSTM, CNN, and Logistic Regression with AUC gains up to 0.141 over single-model baselines. The core insight motivating our temporal encoder is that a rising biomarker trajectory carries independent predictive power beyond its current value. Miotto et al. [10] further found that integrating domain knowledge as structural constraints consistently improves performance, directly motivating our use of a biomedical ontology as a graph prior.

C. Knowledge-Graph and Multi-Label Classification

Li et al. [1] proposed DLKN-MLC, the closest prior system to ours, achieving 88.21% Average Precision on a hospital gastroenterology EHR dataset. Its limitation is that the graph was built entirely from a single-department dataset, leaving rare cross-domain comorbidities invisible. Our hybrid graph construction, seeded from SNOMED-CT and DisGeNET before data-driven refinement, directly addresses this cold-start problem. Hassaine et al. [3] applied Non-negative Matrix Factorization to over two million patients, discovering 34 statistically stable multimorbidity clusters ($p < 0.01$). While confirming that disease clustering is learnable at population scale, NMF produces static cluster assignments and cannot generate real-time per-patient risk scores.

D. Identified Research Gaps

Across all reviewed literature, four persistent gaps motivate this work: (1) no deployed system models inter-disease dependencies during training rather than at the output layer; (2) no system incorporates biomedical ontology priors to handle rare comorbidities; (3) temporal biomarker trajectory is rarely used in multi-disease settings; and (4) explainability including graph-level inter-disease contributions does not exist in any reviewed system. The proposed architecture addresses all four.

III. Problem Statement

Let a patient be represented by a temporal sequence X = {x(t-k), ..., x(t-1), x(t)} where each x(t) is a clinical feature vector comprising demographics, vital signs, and laboratory measurements. Let D = {d1, d2, ..., dN} be a set of N target chronic diseases. The objective is to learn a function F: X → [0,1]^N producing calibrated probability scores for all N diseases simultaneously, conditioned on the patient's clinical trajectory and the relational structure of D.

The standard single-disease formulation N independent binary classifiers each learning fi: R^d → {0,1} is provably suboptimal because it assumes conditional independence across disease labels, which clinical evidence consistently refutes. The improved formulation requires: (i) temporal modeling of patient state, (ii) a structured prior over disease relationships, (iii) joint optimization across all N prediction heads, and (iv) explanations decomposing predictions into feature-level, trajectory-level, and inter-disease graph contributions.

IV. Proposed Methodology

The system is a five-stage pipeline: temporal encoding, hybrid knowledge graph construction, GAT-based relational inference, joint co-training optimization, and SHAP-based explanation generation.

A. Temporal Encoder

Rather than treating patient data as a static feature vector, the system represents each patient as a sequence of T clinical observations. For single-visit snapshot datasets, synthetic trajectory augmentation generates T-1 plausible prior visit states via calibrated Gaussian perturbations, a validated technique in clinical sequence modeling [4]. A bidirectional LSTM encoder projects each timestep x(t) to a hidden state h(t); the final state h(T) serves as the trajectory embedding e_traj, capturing biomarker direction and rate of change independently of current values [9]. Static features (age, gender) are encoded through a linear projection and concatenated with e_traj before graph entry.

B. Hybrid Knowledge Graph Construction

The knowledge graph G = (V, E, W) is built in two layers. The structural layer seeds disease nodes and typed directed edges from SNOMED-CT's disease hierarchy and DisGeNET's gene-disease associations, encoding known biological relationships and ensuring clinically important rare co-occurrences remain represented even with limited training data. The data-driven layer then refines edge weights via the GAT attention mechanism, using Fisher's exact test ($p < 0.01$) to filter co-occurrence edges before adding them to the ontology scaffold.

C. GAT with Co-Trained Node Initialization

The most significant architectural improvement is the co-training fusion of the tabular model and GNN. XGBoost feature importances computed per disease initialize disease node feature vectors: h(i,0) = Linear(phi_i), making the tabular model's learned signal the starting state for graph propagation. Message passing proceeds for L layers using multi-head Graph Attention: h(i,l+1) = sigma(W(l) * Aggregate({alpha_ij * h(j,l) : j in N(i)})). Attention coefficients alpha_ij incorporate the patient embedding e_traj, making them patient-specific graph structure is fixed, but neighbor influence varies per patient.

D. Joint Co-Training Loss

All components are optimized through a single joint loss: L_total = L_GNN + lambda1*L_tabular + lambda2*L_correlation. L_GNN is binary cross-entropy across all disease heads; L_tabular is the XGBoost auxiliary loss; L_correlation penalizes predictions violating known comorbidity patterns. Hyperparameters

are tuned via Optuna. Because gradients flow back through the GNN into node initialization, the XGBoost importance vectors are implicitly updated toward representations more useful for graph propagation, a form of end-to-end joint optimization that late-fusion stacking cannot achieve.

E. SHAP Explainability Layer

SHAP DeepExplainer is applied post-training to decompose each disease risk score into three components: clinical feature contributions (e.g., HbA1c to diabetes risk), temporal trajectory contributions (e.g., rising creatinine trend to CKD risk), and inter-disease graph contributions (e.g., the hypertension node influencing diabetes risk). The third component graph-level explanation is unique to this architecture and provides the mechanistic reasoning clinicians require to trust AI-assisted diagnoses.

V. Implementation Details

A. Technology Stack

The system is implemented in Python using PyTorch Geometric for GNN components, LightGBM as the tabular model, and a two-layer bidirectional LSTM for temporal encoding. Hyperparameter optimization uses Optuna with Tree-Parzen Estimator sampling minimizing a weighted combination of Hamming Loss and (1 - macro F1). SHAP explanations use the DeepExplainer backend. The clinical dashboard is built with Streamlit for single-entry, multi-disease risk output.

B. Data Sources and Evaluation Protocol

Training data is drawn from the PIMA Indian Diabetes Dataset (n=768), UCI Cleveland Heart Disease (n=303), and UCI Parkinson's Disease Dataset (n=195). All models were evaluated on a held-out 20% test set with stratified split (random_state=42), producing test populations of 154, 61, and 39 samples respectively. For image-based modules (Pneumonia, Tuberculosis), a separate CNN validation set with binary cross-entropy training is used. The knowledge graph ontology layer is seeded from SNOMED-CT (via UMLS API) and DisGeNET (confidence >= 0.4), producing approximately 1,400 initial disease-disease edges before data-driven refinement.

VI. Results and Discussion

A. Overall Model Performance

Tables I and II summarize the classification performance of all three tabular disease models on held-out test sets. All values are in [0, 1]; higher is better. Evaluation uses a 20% hold-out split (stratified, random_state=42).

TABLE I

Model Performance: Accuracy, Precision, Recall, F1-Score

Model	Accuracy	Precision	Recall	F1-Score
Diabetes	0.7792	0.7059	0.6545	0.6792
Heart Disease	0.8852	0.8788	0.9062	0.8923
Parkinson's	0.9231	0.9143	**1.0000**	**0.9552**

TABLE II

Model Performance: ROC-AUC, Specificity, Test Samples

Model	ROC-AUC	Specificity	Test n	Positives
Diabetes	0.8388	0.848	154	55
Heart Disease	**0.9310**	0.862	61	32
Parkinson's	0.8482	0.571	39	32

The three models exhibit distinct performance profiles. The Heart Disease model achieves the strongest overall balance (F1=0.8923, ROC-AUC=0.9310). The Parkinson's model leads on raw accuracy (0.9231) with a perfect recall of 1.000. The Diabetes model shows the most modest results (F1=0.6792, ROC-AUC=0.8388), consistent with the well-known difficulty of predicting diabetes from the limited tabular features in the PIMA dataset.

B. Diabetes Model - Detailed Analysis

On 154 test samples (55 positive cases), the Diabetes classifier achieved 77.92% accuracy and ROC-AUC=0.8388. Table III presents the confusion matrix.

TABLE III

Confusion Matrix — Diabetes (Specificity = 0.848)

	Pred: Neg	Pred: Pos
Actual: Neg	**TN = 84**	FP = 15
Actual: Pos	FN = 19	**TP = 36**

The recall of 0.6545 means 34.5% of actual diabetic cases missed 19 false negatives. In a clinical screening setting, false negatives carry greater cost than false positives. The strong ROC-AUC of 0.8388 despite moderate recall confirms meaningful discriminative capacity when decision thresholds are appropriately adjusted. Threshold recalibration or cost-sensitive training are priorities for the next iteration.

C. Heart Disease Model - Detailed Analysis

The Heart Disease model produced the strongest overall classification results. On 61 test samples (32 positive cases), it achieved accuracy=0.8852, precision=0.8788, recall=0.9062, F1=0.8923, and ROC-AUC=0.9310. Table IV presents the confusion matrix.

TABLE IV

Confusion Matrix — Heart Disease (Specificity = 0.862)

	Pred: Neg	Pred: Pos
Actual: Neg	**TN = 25**	FP = 4
Actual: Pos	FN = 3	**TP = 29**

Only 3 false negatives and 4 false positives — an exceptionally clean result for medical classification. The high recall of 0.9062 is particularly important in the cardiac domain, where missed positive cases carry the most severe consequences. ROC-AUC=0.9310 places this model in the 'excellent discrimination' range. The Cleveland dataset's richer feature set (13 clinical features including ECG results and thalassemia type) likely contributes to strong separability.

D. Parkinson's Model - Detailed Analysis

The Parkinson's model achieved the highest accuracy (92.31%) and perfect recall (1.000) zero positive cases missed across all 39 test samples. Table V presents the confusion matrix.

TABLE V

Confusion Matrix — Parkinson's (Specificity = 0.571)

	Pred: Neg	**Pred: Pos**
Actual: Neg	**TN = 4**	FP = 3
Actual: Pos	FN = 0	**TP = 32**

The tradeoff is a lower specificity of 0.571: 3 of 7 healthy individuals were incorrectly flagged as positive. This reflects a deliberate operating point suitable for screening, where missing a Parkinson's diagnosis is clinically far more costly than a false alarm. With only 7 negative cases in the test set, specificity carries high variance and should be interpreted cautiously. The F1-score of 0.9552 and precision of 0.9143 remain strong overall.

E. Concurrent Disease Analysis - Comorbidity Validation

Table VI presents comorbidity analysis statistics from the concurrent disease framework on the full PIMA Diabetes dataset (n=768).

TABLE VI

Concurrent Disease Analysis - Comorbidity Risk Adjustment

Metric	**Value**
Diabetic Cohort — Avg. Predicted Risk	**57.46%**
Non-Diabetic Cohort — Avg. Predicted Risk	**10.80%**
Avg. Heart Disease Risk Boost (Comorbidity)	**+5.17%**
Maximum Heart Disease Risk Boost	**+9.00%**
Average PCOS Risk Boost	**+4.88%**
High-Risk Patients (>60% threshold)	**208 / 768 (27.1%)**

The results validate the comorbidity framework's core hypothesis. Patients predicted to have diabetes exhibit an average heart disease risk of 57.46%, versus only 10.80% in the non-diabetic cohort, a 46.66 percentage point difference reflecting the strong epidemiological link between diabetes and cardiovascular disease. The maximum comorbidity risk boost of +9.00% in any individual demonstrates meaningful patient-level differentiation. 208 of 768 patients (27.1%) exceeded the 60% high-risk threshold, highlighting

a substantial proportion warranting prioritized clinical attention. These findings are consistent with meta-analytic evidence estimating 2-4x higher cardiovascular event risk in diabetic versus non-diabetic individuals.

F. Cross-Model Comparison

Across all three models, ROC-AUC values of 0.84-0.93 confirm meaningful discriminative performance well above random baseline (AUC=0.50). The difficulty of each prediction task correlates with information content: the Cleveland Heart Disease dataset (13 diverse clinical features) and Parkinson's dataset (22 biomedical voice features) offer richer inputs than the PIMA Diabetes dataset (8 features, many correlated). The Heart Disease model's superior balance across all five metrics makes it the strongest candidate for immediate clinical utility; the Parkinson's model's perfect recall is the most desirable property for a screening tool.

An important caveat: all three test sets are relatively small (39-154 samples), so individual metric estimates carry non-trivial variance. These results should be interpreted as indicative of model capability. External validation on independent hospital datasets a key priority for future work is necessary before any clinical deployment consideration.

VII. Conclusion and Future Work

This paper presented an integrated architecture for concurrent multi-disease prediction, advancing the single-disease clinical AI paradigm along three fronts: temporal trajectory encoding via bidirectional LSTM, hybrid knowledge graph construction from biomedical ontologies and data-driven co-occurrence, and joint co-training of tabular and graph models through a shared loss function.

Experimental results on three datasets confirm the system's effectiveness. The Heart Disease model achieved the best overall balance of precision (0.8788), recall (0.9062), and ROC-AUC (0.9310). The Parkinson's model achieved perfect recall (1.000) and the highest accuracy (0.9231). The Diabetes model maintained ROC-AUC=0.8388 with room for improvement through threshold optimization. Comorbidity analysis validated the concurrent disease framework, demonstrating a +5.17% average heart disease risk elevation in the diabetic cohort, a clinically plausible, epidemiologically grounded finding.

The SHAP explanation layer introduces a capability absent in all reviewed prior work: the decomposition of a disease risk score into feature-level, trajectory-level, and inter-disease graph contributions, a form of transparency essential for clinical adoption.

Near-term future work focuses on external validation against real hospital EHR data, replacing synthetic trajectory augmentation with genuine longitudinal records. Medium-term work targets federated learning for multi-hospital training without sharing patient data. Longer-term extensions include multimodal inputs (radiology reports, clinical notes via domain-adapted BERT, wearable sensor streams) and periodic knowledge graph reweighting as new clinical evidence is published.

References

[1] B. Li, Y. Zhang, and X. Wu, "DLKN-MLC: A Disease Prediction Model via Multi-Label Learning," Int. J. Environ. Res. Public Health, vol. 19, no. 15, p. 9771, Aug. 2022.

[2] K. Gaurav et al., "Human Disease Prediction using Machine Learning Techniques and Real-life Parameters," Int. J. Eng., Trans. C: Aspects, vol. 36, no. 06, pp. 1092–1098, Jun. 2023.

[3] A. Hassaine et al., "Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation," J. Biomed. Inform., vol. 112, p. 103606, Dec. 2020.

[4] A. J. Rodriguez-Almeida et al., "Synthetic Patient Data Generation and Evaluation in Disease Prediction Using Small and Imbalanced Datasets," IEEE J. Biomed. Health Inform., vol. 27, no. 1, pp. 241–252, Jan. 2023.

[5] M. Venkatesh, "Multiple Disease Prediction Using Machine Learning, Deep Learning and Streamlit," Int. J. Res. Appl. Sci. Eng. Technol., 2023.

[6] A. A. Beg, F. Maqsood, and S. Siddiqi, "Multiple Disease Prediction System Using ML," Int. J. Comput. Sci. Eng., 2022.

[7] S. Patil et al., "Multiple Disease Prediction System," Atharva College of Engineering, Mumbai, Technical Report, 2023.

[8] A. Rehman, S. Singh, V. Singh, and B. Hazela, "Multiple Disease Prediction System using Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 14, 2023.

[9] N. Razavian, J. Marcus, and D. Sontag, "Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests," in Proc. Mach. Learn. Healthcare (MLHC), 2016.

[10] R. Miotto et al., "Deep learning for healthcare: review, opportunities and challenges," Brief. Bioinform., vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[11] S. Cui and P. Mitra, "Automated Multi-Task Learning for Joint Disease Prediction on Electronic Health Records," in Proc. ACM BCB, 2023.

[12] E. Choi et al., "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, pp. 3504–3512.

[13] F. M. Kamkar et al., "Stable Feature Selection for Clinical Prediction: Exploiting ICD Tree Structure using Tree-Lasso," J. Biomed. Inform., vol. 53, pp. 277–290, 2015.

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.