Understanding Structured Health Data through Interaction-Aware Mixture-of-Experts

Anonymous Authors[1]

Abstract

We study interaction-aware mixture-of-experts for post-stroke rigidity prediction using multi-level views of structured health records. Despite minimal performance gains, routing attribution reveals systematic importance differences across views, underscoring view construction as key to interpretability.

1. Introduction

Structured health data are a dominant substrate for clinical prediction, spanning billing codes, diagnoses, procedures, vital signs, and other irregular clinical measurements organized in tabular form (Shickel et al., 2018; Xu et al., 2025). Common modeling approaches reflect this structure, as gradient-boosted decision trees and tabular neural networks remain strong baselines on structured health-data tasks (Chen & Guestrin, 2016; Gorishniy et al., 2021; Somepalli et al., 2021; Popov et al., 2020; Wang et al., 2021; Shwartz-Ziv & Armon, 2022; McElfresh et al., 2023). Yet most structured-data models either treat the record as a single tabular input, leaving clinically meaningful interactions to be learned implicitly, or attach naturally distinct modalities such as text or imaging. It remains unclear whether transforming a single structured record into multiple alternative representations, and modeling interactions among them, can improve predictive performance while making the prediction process easier to interpret.

To explore this question, we adopt the terminology of multi-view learning, where a *view* denotes an alternative representation of the same underlying example (Sun, 2013). Multi-view learning is designed to exploit complementary and shared information across views, and has been applied broadly—for instance, to fuse imaging and clinical text in medical diagnosis (Wang et al., 2018) or to combine different sensor modalities in activity recognition (Zhang et al., 2019). How to construct and exploit multiple views of a *single* structured health record, however, remains underexplored.

For structured health data, such views can be constructed at multiple levels. At the *model level*, when the same structured input is passed through different predictive models, each model's learned representation constitutes a view, since different architectures encode different inductive biases (Gorishniy et al., 2021; Popov et al., 2020; Shwartz-Ziv & Armon, 2022). At the *data level*, the same structured input can be partitioned by clinical semantics into administrative, procedural, diagnostic, vital-sign, and code groups (Johnson et al., 2023; Ma et al., 2025). At the *representation level*, the record can be encoded through different paradigms—graph-based patient–feature structure (Brody et al., 2022; Choi et al., 2020; Rocheteau et al., 2021), tabular representation learning, and text embeddings from natural-language renderings of the record (Hegselmann et al., 2023; Lee et al., 2024; Steinberg et al., 2021). Furthermore, While simple concatenation of different views has the potential to improve predictive performance, it does not distinguish view-specific signal from information shared across views or emerging only through their combination, which can be critical for providing explanation in sensitive contexts such as healthcare.

To characterize these distinctions principally, we employ the Partial Information Decomposition (PID) framework, which decomposes the information that a set of views carries about a target into redundancy, uniqueness, and synergy (Williams & Beer, 2010; Bertschinger et al., 2014; Liang et al., 2023). Specifically, we adopt the recently proposed I2MoE framework (Xin et al., 2025), which routes inputs through specialized experts and explicitly present view-specific and synergistic signals, to study post-stroke rigidity prediction using a national inpatient stroke cohort derived 129,401 hospital admissions. Using this setup, we examine whether model-level, data-level, and representation-level views—all derived from the same structured health record—provide useful interaction structure for prediction and interpretation.

Our results show that introducing a multi-view approach yields minimal improvements in predictive performance,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

consistent with views being alternative decompositions of the same record rather than independent modalities. Nevertheless, routing attribution reveals that model-level, data-level, and representation-level view designs produce systematically different allocations of importance across distinct views, and that these allocations are locally consistent across similar patients. These findings suggest that view construction is a meaningful design choice for interpretability even when predictive gains are modest, and point toward routing-based attribution as a practical lens for building explanations from structured health records.

2. Related Work

Structured tabular health modeling. Structured health data are central to clinical prediction because they capture diagnoses, procedures, utilization, and longitudinal measurements at scale (Shickel et al., 2018; Xu et al., 2025). These data are commonly modeled as a single tabular representation. Gradient-boosted trees, including XGBoost, remain strong baselines for tabular prediction (Chen & Guestrin, 2016), while tabular neural architectures such as SAINT, NODE, and DCN-V2 provide competitive alternatives for structured inputs (Somepalli et al., 2021; Popov et al., 2020; Wang et al., 2021). These methods, however, largely preserve a single-view treatment of the record and abstract away from the distinct clinical concepts each measure represent. We ask whether a single structured health record can instead be expressed through multiple view definitions, and whether those views make interaction-aware prediction and interpretation more informative.

Interaction modeling and interpretability. Clinical prediction from structured records often depends on how variables act together rather than in isolation. Interpretable additive models with pairwise terms have shown that selected feature interactions can support clinical risk prediction while remaining inspectable (Caruana et al., 2015), while neural interaction detection and post-hoc attribution methods explain dependencies in trained predictors (Tsang et al., 2018; Lundberg & Lee, 2017). These approaches primarily operate at the level of variables or variable pairs. In contrast, view-based modeling organizes a structured record into higher-level sources before fusion, shifting the question to whether each view contributes distinct, duplicated, or jointly useful signal. Partial Information Decomposition formalizes these cases as unique, redundant, and synergistic information (Williams & Beer, 2010; Bertschinger et al., 2014; Liang et al., 2023). I2MoE implements this idea in a predictive MoE through unique, synergy, and redundancy experts (Xin et al., 2025). Our work uses this framework to evaluate whether view-level interaction modeling is useful when all views are constructed from the same structured health record.

3. Methods

3.1. Task Definition

To study how interaction-aware modeling behaves when a single structured record is expressed through alternative views, we use post-stroke rigidity prediction as a binary classification task. The cohort is derived from adult HCUP/NIS stroke hospitalizations from 2016–2020, consisting of 129,401 admissions. HCUP/NIS is an all-payer database of U.S. hospital inpatient stays derived from hospital billing data and includes clinical and resource-use information typically available from discharge abstracts. We define rigidity using a clinician-curated set of 55 ICD-10-CM codes. Selected predictor variables span administrative factors, procedures, diagnoses, clinical signs, and grouped ICD indicators.

3.2. Three View Definitions

For structured health data, the view structure is often a modeling choice rather than a fixed property of the raw record. The structured nature of these records allows alternative views to be defined at different stages of the modeling pipeline: before modeling by grouping variables, during modeling through architecture-specific encoders, and after upstream encoding by fusing learned representations. We use these three stages to define data-level, model-level, and representation-level views while keeping the I2MoE formulation fixed.

3.2.1. Model-Level Views

The model-level formulation treats representations from tabular prediction backbones as views. We encode the full processed record with SAINT, NODE, and DCN-V2 (Somepalli et al., 2021; Popov et al., 2020; Wang et al., 2021), three models designed for structured tabular inputs with different inductive biases: attention-based feature modeling, tree-inspired representation learning, and explicit feature crossing. Their hidden representations are passed to separate unique experts, evaluating whether tabular-specialized models produce distinct predictive signals from the same record.

3.2.2. Data-Level Views

The data-level formulation defines views by semantic feature partition. The record is split into administrative, procedure, diagnosis, clinical sign, and ICD-indicator groups. Each group is encoded by a separate multilayer perceptron unique expert. This examines whether interaction structure emerges when views correspond to clinically meaningful subdomains of the same structured record.

Table 1. Summary of predictive performance. * denotes a statistically significant improvement over XGBoost across random seeds.

Model	Type	Setting	AUROC	AUPRC	F1
XGBoost	Machine Learning	–	0.7627	0.7786	0.7182
GATv2	Graph-Based	–	0.7512	0.7578	0.7235
SAINT	Tabular DL	–	0.7625	0.7789	0.7170
NODE	Tabular DL	–	0.7634	0.7799	0.7160
DCN-V2	Tabular DL	–	0.7626	0.7789	0.7154
I^2MoE (Model-Level)	Mixture of Experts	Full	0.7714*	0.7888*	0.7213
		No Synergy	0.7713*	0.7887*	0.7223
		No Redundancy	0.7713*	0.7888*	0.7226
		No S & R	0.7713*	0.7886*	0.7171
I^2MoE (Data-Level)	Mixture of Experts	Full	0.7621	0.7788	0.7182
		No Synergy	0.7621	0.7787	0.7185
		No Redundancy	0.7621	0.7788	0.7186
		No S & R	0.7622	0.7786	0.7178
I^2MoE (Representation-Level)	Mixture of Experts	Full	0.7600	0.7781	0.7313*
		No Synergy	0.7600	0.7781	0.7317*
		No Redundancy	0.7601	0.7781	0.7307*
		No S & R	0.7601	0.7781	0.7316*

3.2.3. Representation-Level Views

The representation-level formulation defines views after upstream encoding. For each patient, we use a frozen GATv2 encoder (Brody et al., 2022) to obtain a graph representation, a frozen NODE encoder (Popov et al., 2020) to obtain a tabular representation, and a frozen Qwen3-Embedding encoder (Zhang et al., 2025) applied to a structured-text rendering of the record to obtain a text representation. These embeddings are then fused by I2MoE. This evaluates whether interaction-aware fusion is more useful after view-specific representation learning.

3.3. I2MoE Formulation and Objective

Let $\{\mathbf{h}_m\}_{m=1}^{M}$ denote the view-specific representations, with $M{=}3$ for model-level and representation-level views and $M{=}5$ for data-level views. Following I2MoE (Xin et al., 2025), we use one unique expert per view, one synergy expert, one redundancy expert, and a reweighting network g:

$$\hat{y} \;=\; \sigma\left(\sum_{m=1}^{M} w_m^u f_m^u(\mathbf{h}_m) + w^s f^s(\mathbf{h}) + w^r f^r(\mathbf{h})\right), \tag{1}$$

where $[w_1^u, \ldots, w_M^u, w^s, w^r] = g([\mathbf{h}_1; \ldots; \mathbf{h}_M])$. The synergy expert is intended to capture signal that emerges only when views are considered jointly, whereas the redundancy expert captures signal shared across views.

We optimize the task loss and the weakly supervised I2MoE interaction loss, which encourages unique, synergy, and redundancy experts to capture view-specific, combined, and shared information:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{task}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}}, \tag{2}$$

We refer readers to Xin et al. (2025) for the full construction of $\mathcal{L}_{\text{int}}$.

3.4. Experimental Protocol

For each view definition, we train a full I2MoE model with unique, synergy, and redundancy experts, together with three ablations that remove the synergy expert, the redundancy expert, or both. We evaluate predictive performance using AUROC, AUPRC, and F1, with all predictive results averaged over 30 random seeds. We use expert-routing weights for interpretation at two levels. Global interpretation summarizes how each view definition allocates predictive mass across unique, synergy, and redundancy experts at the cohort level, estimated by averaging expert-routing weights over the test cohort and across retrainings. Local interpretation asks whether patients with similar learned representations receive similar expert-weight allocations. For this, we use the representation-level model and compare expert-routing distances between nearest-neighbor patients in the learned representation space and randomly matched patients.

4. Experiments and Results

4.1. Predictive Performance

The three view definitions are competitive with strong graph and tabular baselines, but their advantages are metric-dependent (Table 1). Compared with XGBoost, the model-level view shows a statistically significant increase in AUROC and AUPRC, while the representation-level view shows a statistically significant increase in F1 but lower AUROC. The data-level view is not significantly different from XGBoost on the main metrics. This is the central

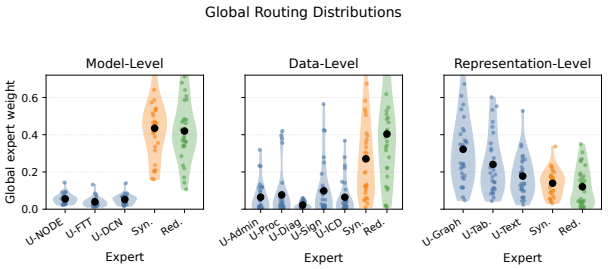

Figure 1. Cohort-level expert routing across retrainings.

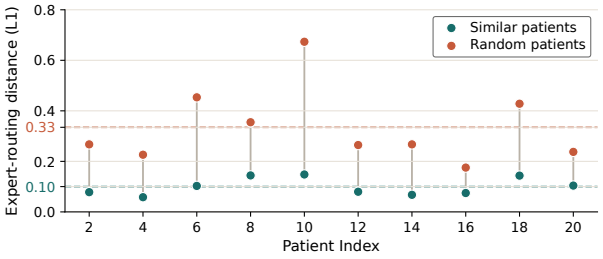

Figure 2. Patient-level routing consistency.

empirical pattern of our study. A single structured health record can be re-expressed as model-level, data-level, or representation-level views, and these choices change prediction behavior. However, the performance gains are limited because the views are alternative decompositions of the same underlying record rather than independent modalities.

4.2. Interaction Ablation

The interaction ablations in Table 1 show no statistically significant degradation after removing synergy, redundancy, or both. This suggests that the interaction experts are not the main source of the observed predictive performance. Instead, most predictive signal appears recoverable from the unique experts and the reweighting network. In this setting, synergy and redundancy experts are more diagnostic than performance-improving: they expose how the model allocates mass to view-specific, shared, and combined signals, but do not by themselves produce a clear predictive gain. This suggests that, when views are constructed from the same structured record rather than separate data sources, extracting additional predictive signal may require stronger interaction modeling or more distinctive view construction.

4.3. Global Interpretation

I2MoE provides routing-based interpretation through a reweighting network that assigns an expert weight to each unique, synergy, and redundancy expert for each prediction. Averaging these weights over the test cohort summarizes how the model allocates prediction across the view-specific, shared, and jointly useful signals defined in Section 3.3 (Figure 1). Model-level views place high mass on both synergy and redundancy experts. This is consistent with the fact that the views are different tabular backbones trained on the same input, whose learned representations can overlap substantially while still producing complementary decision patterns. Data-level views also route substantial mass through synergy and redundancy experts, reflecting the dependence among clinical groups such as diagnoses, procedures, signs, and administrative factors. These groups are clinically distinct but not statistically independent, so shared and combined signal are expected. In contrast, representation-level views allocate more mass to unique experts, suggesting that graph, tabular, and text-derived embeddings preserve more view-specific signal. Overall, global expert weights show that view definition changes how the model distributes prediction across view-specific and interaction experts. We interpret these weights as model allocation patterns rather than direct clinical evidence of synergy or redundancy, since all views are derived from the same structured record.

4.4. Local Interpretation

We further study how the model assigns expert weights for individual patients through the I2MoE reweighting network. This analysis is performed for the representation-level model, where each patient has a learned graph-tabular-text representation. To study patient-neighborhood behavior, we select 10 evaluation patients and use each one to define a local neighborhood in this representation space. For each selected patient, we compare the expert-weight distributions of its nearest neighbors with those of randomly matched patients (Figure 2). Nearby patients have consistently smaller expert-weight distance than random controls, suggesting that patient-level expert weights vary in a locally consistent way rather than changing arbitrarily across similar cases. Such local consistency is an encouraging property that enables future work on utilizing expert routing information to provide meaningful patient-level explanations.

5. Conclusion

We examined interaction-aware mixture-of-experts modeling for post-stroke rigidity prediction when one structured health record is expressed through model-level, data-level, and representation-level views. Multi-view modeling yields metric-dependent but limited predictive gains, and removing interaction experts does not significantly degrade performance—suggesting that explicit synergy and redundancy modeling is more useful for diagnosing how views are used than for improving prediction. Routing analyses reveal that view construction choices produce systematically different expert allocations at the cohort level, and that these allocations are locally consistent across similar patients. Together, these findings position routing-based attribution as a promising foundation for offering accurate prediction with patient-level explanation in healthcare context based on structured data. Future work in this direction can aim to connect expert-level routing to finer-grained clinical concepts, and developing view constructions that extract more distinctive signal from the same underlying record.

References

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? In *ICLR*, 2022.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, pp. 1721–1730, 2015.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *KDD*, pp. 785–794, 2016.

Choi, E., Xu, Z., Li, Y., Dusenberry, M. W., Flores, G., Xue, Y., and Dai, A. M. Learning the graphical structure of electronic health records with graph convolutional transformer. In *AAAI*, 2020.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. In *NeurIPS*, 2021.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. TabLLM: Few-shot classification of tabular data with large language models. In *AISTATS*, 2023.

Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Lee, S. A. et al. Multimodal clinical pseudo-notes for emergency department prediction tasks using multiple embedding model for EHR (MEME). *arXiv preprint arXiv:2402.00160*, 2024.

Liang, P. P., Cheng, Y., Fan, X., Ling, C. K., Nie, S., Chen, R. J., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., and Morency, L.-P. Quantifying & modeling multimodal interactions: An information decomposition framework. In *NeurIPS*, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Ma, L. et al. Harnessing the potential of multimodal EHR data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion*, 2025.

McElfresh, D. et al. When do neural nets outperform boosted trees on tabular data? In *NeurIPS Datasets and Benchmarks*, 2023.

Popov, S., Morozov, S., and Babenko, A. Neural oblivious decision ensembles for deep learning on tabular data. In *ICLR*, 2020.

Rocheteau, E., Tong, C., Veličković, P., Lane, N., and Liò, P. Predicting patient outcomes with graph representation learning. In *AAAI W3PHIAI Workshop*, 2021.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5): 1589–1604, 2018.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., and Goldstein, T. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., and Shah, N. H. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113: 103637, 2021.

Sun, S. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7–8):2031–2038, 2013.

Tsang, M., Cheng, D., and Liu, Y. Detecting statistical interactions from neural network weights. In *ICLR*, 2018.

Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., and Chi, E. DCN V2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *The Web Conference*, pp. 1785–1797, 2021.

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058, 2018.

Williams, P. L. and Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Xin, J., Yun, S., Peng, J., Choi, I., Ballard, J. L., Chen, T., and Long, Q. I2MoE: Interpretable multimodal interaction-aware mixture-of-experts. In *ICML*, 2025.

Xu, W. et al. A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models. *arXiv preprint arXiv:2507.12774*, 2025.

Zhang, X., Wong, Y., Kankanhalli, M. S., and Geng, W. Hierarchical multi-view aggregation network for sensor-based human activity recognition. *PLOS ONE*, 14(9): e0221390, 2019. doi: 10.1371/journal.pone.0221390.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.