Trust Your Neighbors: Multimodal Patient Retrieval for TBI Prognosis

Pranav Manjunath, Brian Lerner, and Timothy W. Dunn

***Abstract*— Early and accurate triage of traumatic brain injury is critical for guiding treatment decisions that optimize patient outcomes. A major early clinical decision point occurs in the emergency department, where providers must decide whether to admit or discharge patients with head injuries, yet these decisions are often inconsistent and rarely supported by case-based frameworks. Here, we introduce RAPID-TBI (Retrieval Augmented Prediction for Informed Disposition in Traumatic Brain Injury), a multimodal system that predicts emergency department disposition using example-based retrieval to emulate clinical case-based reasoning. RAPID-TBI achieves state-of-the-art classification performance while enhancing interpretability by retrieving similar patients to inform predictions. Using a large multimodal TBI dataset from a major U.S. hospital system, RAPID-TBI integrates head CT scans, radiology reports, exam findings, laboratory values, vitals, and demographics through an attention-based encoder that generates patient embeddings for disposition classification. We further assessed RAPID-TBI across institutional and temporal generalizability, showing consistent performance and resilience to shifts in data distribution. Finally, we explored small language models as prompt-based classifiers for retrieval-guided prediction without fine-tuning. Together, these components enable RAPID-TBI to deliver consistent, individualized, and clinically grounded predictions, a promising step toward trustworthy, personalized decision support in TBI care.**

***Index Terms*— Multimodal AI, Example-based XAI, Traumatic Brain Injury**

I. Introduction

TRAUMATIC brain injury (TBI) is a significant health challenge, requiring complex clinical management due to its heterogeneous nature, with early and accurate prognosis playing a crucial role in guiding treatment decisions [1], [2]. Care of patients with TBI, who are often first evaluated in the emergency department (ED), generates vast amounts of multimodal data, including CT imaging, radiology reports, clinical notes, and structured measurements such as vitals, exams, lab, and administered medicines. These diverse data sources provide complementary insights for patient assessment, triage, and prognosis. A major concern with clinical management is the significant variability in how TBI patients are discharged home or admitted for further observation, driven by a lack of strong evidence supporting these decisions [3]. This suggests that many inpatient admissions may be unnecessary, leading to prolonged hospital stays, increased health risks, patient burden, and excessive healthcare costs [4]. By developing predictive models for ED disposition, we can enhance clinical decision-making to improve patient outcomes and reduce socio-economic burden.

Historically, TBI prognosis has relied on risk scores derived from structured clinical variables [5], [6]. However, recent advances in deep learning and increased availability of EHR data have enabled the integration of multimodal inputs, including imaging, clinical notes, and structured features, leading to more powerful and nuanced prognostic models for TBI care [7]–[11]. Importantly, ED disposition decisions require synthesizing such multimodal information across multiple time points to monitor changes in patient status during their ED stay [12]. However, most existing TBI datasets lack this temporal and comprehensive multimodal dimension. To address this gap, we curate a comprehensive multimodal TBI dataset that captures longitudinal measurements throughout the ED stay and spans the full spectrum of injury severity.

To develop effective models for advancing TBI care, it is essential to account for the inherent variability in patient presentations, enabling personalized approaches that move beyond population averages toward individualized, data-driven decisions informed by clinical, imaging, and other multimodal factors [13]–[15]. In practice, clinicians routinely synthesize vast amounts of clinical data to guide care, often recalling similar past cases as precedent [16]. We envision a clinician-assistive system that emulates real-world clinical reasoning by integrating multimodal patient data, retrieving relevant prior cases, and reasoning about them transparently to support informed decisions. Such a system would not only deliver accurate predictions but also provide transparency through retrieved examples, enabling clinicians to better understand, validate, and act on model outputs [17].

To support this vision, example-based explainable AI (XAI), synonymous to case-based reasoning, has emerged as a promising approach. By grounding predictions in examples from similar patients, it offers a more interpretable alternative to black-box models and supports individualized, context-aware decision-making. Retrieval systems are central to this approach, enabling the rapid and accurate identification of relevant clinical examples that underpin personalized care [18]. This positions them as key components in advancing precision health [19]. This paradigm has gained traction in clinical AI

Pranav Manjunath is with the Department of Biomedical Engineering at Duke University, Durham, NC, USA (e-mail:pranav.manjunath@duke.edu)

Brian Lerner is with the Department of Electrical and Computer Engineering at Duke University, Durham, NC, USA (e-mail:brian.lerner@duke.edu)

Timothy W. Dunn is with the Department of Biomedical Engineering at Duke University, Durham, NC, USA (e-mail:timothy.dunn@duke.edu)

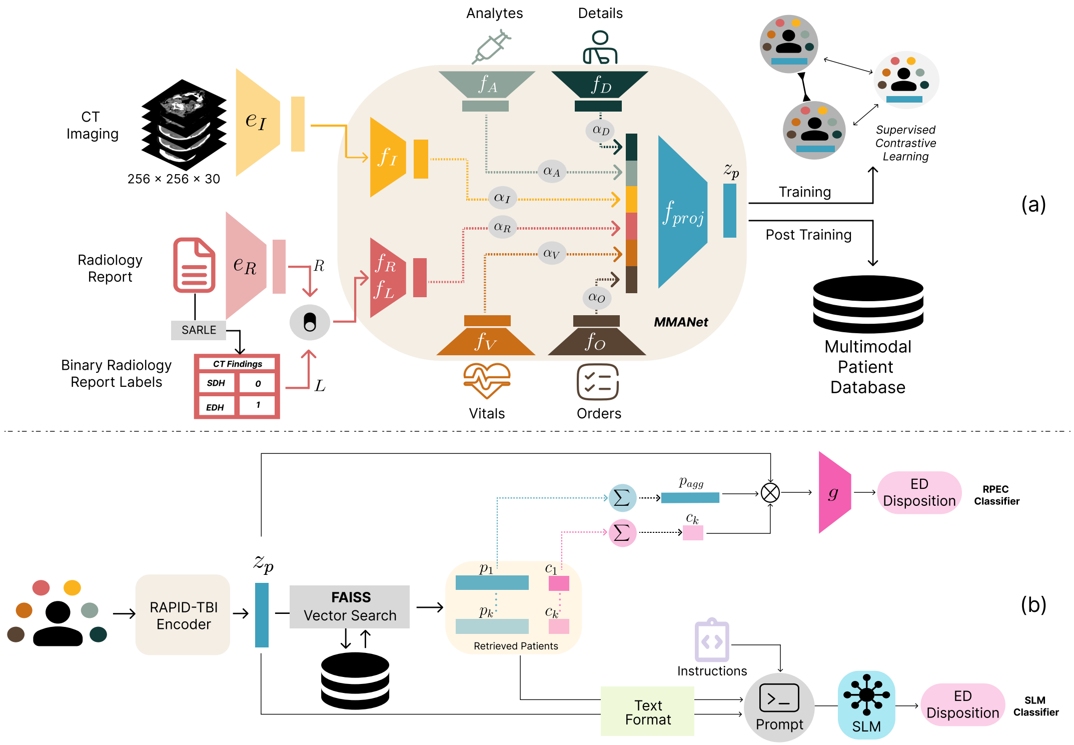

Fig. 1. An overview of RAPID-TBI, consisting of (a) Encoder: Creation of multimodal patient embeddings (z_p) and training using Supervised Contrastive Learning. Post Training: Encoding each patient as a multimodal embedding and storing the embeddings into a database, and (b) Retriever and Classifier: Upon retrieving similar patients to a test patient, top depicts the RPEC classifier while bottom depicts the SLM classifier.

research [20], [21] and has been shown to improve clinician trust [22]. Researchers have explored neighborhood-based classifiers and Graph Neural Networks (GNNs) to enable such reasoning; notably, [23] demonstrated that GraphSAGE effectively models patient similarity networks in the ED setting, linking predictions to prior cases. While such methods achieve comparable accuracy to traditional models [24], [25], few have extended this approach to multimodal clinical data, particularly combinations of structured and un-structured variables.

At the same time, language models have emerged as powerful tools for clinical reasoning. Their ability to generate free-text explanations aligns naturally with how clinicians interpret and communicate decisions [17]. Large language models (LLMs) like GPT-4 have demonstrated strong performance across medical prediction and summarization tasks [26]. However, their reliance on cloud infrastructure, high computational demands, and risks to protected health information limit their use in clinical settings [27]. In contrast, small language models (SLMs) are lightweight, open-source, and deployable on-premise, making them a safer and more practical choice for real-time clinical decision support. However, their smaller capacity can limit reasoning depth and generalization, particularly in complex or less structured clinical scenarios.

As a step toward this goal, we present RAPID-TBI (Retrieval-Augmented Prediction for Informed Disposition in TBI), a personalized, multimodal framework for predicting ED disposition through example-based explanations (Fig. 1). RAPID-TBI comprises three key components: (i) an Encoder, a novel multimodal attention-based encoder (MMANet) (Fig. 1a) that fuses head CT scans, radiology reports, clinical orders, lab results, vital signs, and demographics into unified patient embeddings; (ii) a Retriever (Fig. 1b), which leverages FAISS vector search [28] at inference time to identify similar patients via cosine similarity; and (iii) a Classifier (Fig. 1b), which incorporates the retrieved patients as personalized context to predict ED disposition. We explore two classifier designs. The first, Retrieved Patient-Embedding Fused Classifier (RPEC), is a novel multilayer perceptron (MLP)-based classifier that integrates the embeddings and ED dispositions of retrieved patients with the test patient's embedding to make predictions. The second, leverages open-source SLMs as a classifier through in-context learning. In this setup, we provide the SLM with textual representations of the retrieved similar patients and their ED dispositions as few-shot examples, along with the test patient's texual representation, to predict the ED disposition. The SLM requires no additional training, making it lightweight, adaptable, and deployable in real-world clinical settings. By combining multimodal data integration with example-driven prompting, RAPID-TBI facilitates individualized decision-making, demonstrating how retrieval-augmented prediction can advance precision health. To our knowledge, this represents the first multimodal system that emulates case-based reasoning to support TBI care. Our key contributions:

1) Developed RAPID-TBI, an example-based XAI system for predicting ED disposition in TBI, mirroring the process of personalized case-based reasoning through the retrieval of similar patients.
2) Designed MMANet, an attention-based encoder that

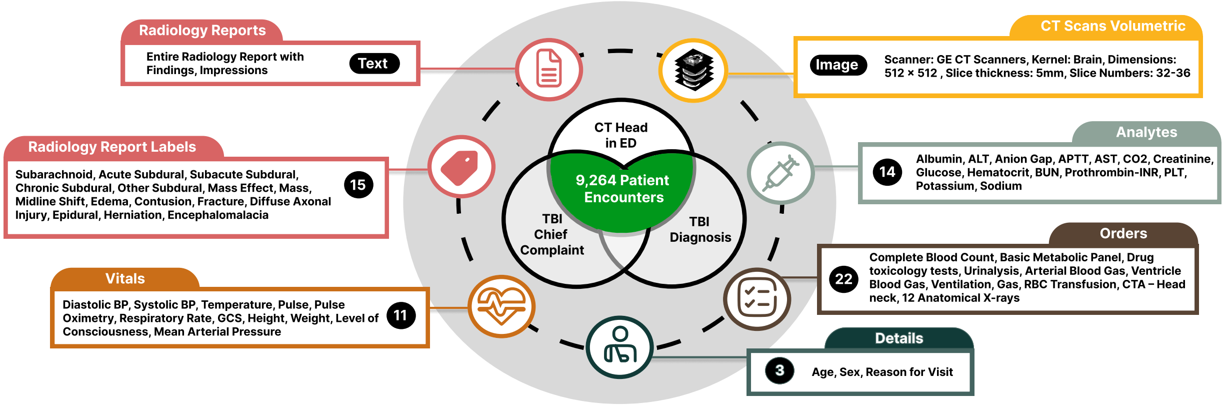
Radiology Reports
Entire Radiology Report with Findings, Impressions
Text
CT Scans Volumetric
Image
Scanner: GE CT Scanners, Kernel: Brain, Dimensions: 512 × 512 , Slice thickness: 5mm, Slice Numbers: 32-36
Radiology Report Labels
Subarachnoid, Acute Subdural, Subacute Subdural, Chronic Subdural, Other Subdural, Mass Effect, Mass, Midline Shift, Edema, Contusion, Fracture, Diffuse Axonal Injury, Epidural, Herniation, Encephalomalacia
15
CT Head in ED
9,264 Patient Encounters
TBI Chief Complaint
TBI Diagnosis
Analytes
14
Albumin, ALT, Anion Gap, APTT, AST, CO2, Creatinine, Glucose, Hematocrit, BUN, Prothrombin-INR, PLT, Potassium, Sodium
Orders
22
Complete Blood Count, Basic Metabolic Panel, Drug toxicology tests, Urinalysis, Arterial Blood Gas, Ventricle Blood Gas, Ventilation, Gas, RBC Transfusion, CTA – Head neck, 12 Anatomical X-rays
Vitals
Diastolic BP, Systolic BP, Temperature, Pulse, Pulse Oximetry, Respiratory Rate, GCS, Height, Weight, Level of Consciousness, Mean Arterial Pressure
11
Details
3
Age, Sex, Reason for Visit

Fig. 2. The center Venn diagram representing cohort criteria. Surrounding boxes represent the multimodal data, where black circles indicate either the data type or the number of tabular features.

integrates longitudinal multimodal ED data via contrastive learning to generate unified patient embeddings, with personalized modality-adaptive attention weights for each patient.

3) Proposed RPEC, a MLP-based classifier that combines embeddings of the test patient and similar retrieved patients to predict ED disposition. RPEC outperforms state-of-the-art black-box multimodal and graph-based models across multiple evaluation metrics.
4) Demonstrated that incorporating retrieved similar patients into prompts through in-context learning for language models improves classification performance.
5) Evaluated generalizability through cross-institutional and temporal experiments, and assessed fairness by examining performance across demographic subgroups.

II. Data

For this study, we curated a retrospective multimodal dataset of electronic health records (EHR) from adult patients with TBI (Fig. 2), including encounters in which patients received a head CT in the ED and had either a TBI-related chief complaint, a documented TBI diagnosis, or both. The cohort was derived from visits to two hospitals within the Duke University Health system between 2015 and 2020. In total, the dataset comprises 9,264 independent patient encounters representing 8,595 unique patients. Each patient encounter is treated as a distinct patient, as trauma-related decisions are encounter-specific and influenced by the unique clinical presentation at each visit [29]. ED disposition in our dataset was initially categorized into four classes: discharge home, floor admission, ICU, and surgery. Due to class imbalance, particularly in the ICU and surgery groups, we binarized the labels for modeling: discharge home (72.4%) vs. hospital admission (27.6%), where the latter combines floor, ICU, and surgery cases. This formulation also enabled exploration of the model's ability to capture latent clinical variables, explained in the Methods section.

Each encounter includes data collected entirely within the ED, encompassing the initial volumetric head CT scan (I), its corresponding radiology report (R), and structured variables such as patient details (D), vital signs (V), laboratory analytes (A), and clinical orders (O). Additionally, we utilized SARLE [30], tuned to head CT findings, to extract binary label findings (L) from the radiology reports. As the vitals and analytes are recorded multiple times throughout the ED course, we summarized them using the arithmetic mean, measurement count (number of times recorded), and range (minimum, maximum), these statistics served as features for model input. Clinical orders are encoded as binary indicators denoting whether a given order was placed. Importantly, the only source of missingness in the dataset came from structured variables, such as vitals and analytes that were not recorded for certain patients. This missingness, which ranges from 0% to 60% across variables, reflects real-world EHR data where not all clinical measurements are consistently documented. The data was stratified by ED disposition into training, validation, and test splits containing 56%, 18%, and 26% of the cohort, respectively. Across the train, validation, and test sets, the proportion of female patients was 56.3%, 57.7%, and 54.9%, respectively. The mean age with standard deviation was consistent across splits (train: 58.66 ± 21.24, val: 58.64 ± 21.13, test: 57.57 ± 21.55 years). Among encounters with non-missing GCS values (roughly 60% across splits), the distribution of TBI severity was stable, with approximately 93% mild, 4–5% moderate, and 2% severe cases in each split.

III. Methods

A. RAPID-TBI Encoder - MMANet

For each encounter, the raw data from each modality $m \in \{I, R, A, O, V, D, L\}$ are represented as x_m. While image and report are distinct modalities, the remaining variables are structured tabular inputs. All CT scans are non-contrast head CTs acquired with a uniform slice thickness of 5 mm. Each volume is preprocessed by clipping voxel intensities to the range $[-1024, 2048]$ HU to remove outliers and normalize the dynamic range, followed by min–max normalization to $[0, 1]$.

To obtain semantically rich representations for the image and text modalities, we first leverage pretrained domain-specific encoders. For volumetric imaging, we use CTNet [30] as the image encoder e_I, which maps a head CT volume $x_I \in \mathbb{R}^{256\times256\times30}$ to a 512-dimensional embedding $e_I(x_I) \in \mathbb{R}^{512}$.

CTNet is pretrained using InfoNCE contrastive learning [31] and remains frozen during multimodal training. Radiology reports x_R are encoded using BioClinicalBERT [32], a language model pretrained on clinical corpora, yielding embeddings $e_R(x_R) \in \mathbb{R}^{728}$. These pretrained encoders are used solely to initialize high-level representations for imaging and text data and are not updated during multimodal training.

The RAPID-TBI multimodal encoder, MMANet, processes all modalities through dedicated modality-specific encoders f_m to produce unified embeddings $z_m = f_m(x_m)$. For imaging and text modalities, the input to dedicated MLP-based encoders (f_I and f_R) are the pretrained representations $e_I(x_I)$ and $e_R(x_R)$, respectively (Eq. 1). For the tabular modalities, the raw features x_m are processed by modality-specific NeuMiss network [33] encoders f_m (Eq. 2), designed to handle missing and heterogeneous data. The resulting embeddings z_m are then fused through an attention-based multimodal fusion layer and optimized using a supervised contrastive learning objective [34].

$$z_m = f_m(e_m(x_m)), \quad \forall m \in \{\mathrm{I}, \mathrm{R}\}, \quad z_m \in \mathbb{R}^{128} \tag{1}$$

$$z_m = f_m(x_m), \quad \forall m \in \{A, V, O, D, L\}, \quad z_m \in \mathbb{R}^{128} \tag{2}$$

$$\alpha_m = \frac{\exp(f_m^{\text{att}}(z_m))}{\displaystyle\sum_{b \in \{I,R,A,O,V,D\}} \exp(f_b^{\text{att}}(z_b))} \tag{3}$$
$$\text{for all } m \in \{I, R, A, O, V, D\}$$

$$z_p = f_{\text{proj}}\left(\bigotimes_{m \in \{I,R,A,O,V,D\}} (\alpha_m \cdot z_m) \right), \quad z_p \in \mathbb{R}^{128} \tag{4}$$

Modality-specific attention layers f_m^{att} compute raw attention modality scores, which are normalized via softmax to obtain α_m (Eq.3). These weights are applied to z_m to form weighted modality embeddings, which are concatenated (concatenation operator represented as $\bigotimes$) and transformed through a projection layer (Eq. 4) into the final patient embedding z_p, which is then normalized on the hypersphere. During contrastive learning, patient pairs with the same ED disposition are treated as positives. The multimodal architecture receives either x_R or x_L as input, but not both together, to assess the effectiveness of radiology report representations in ED disposition prediction.

In addition, we examine how clinically relevant variables, both explicit features and latent attributes not directly provided during training, are represented within the multimodal patient embeddings. We conduct three experiments: (i) understanding deviation of age (an explicit variable) between patient and retrieved patients, (ii) assessing whether report labels were latently encoded for models using radiology reports, and (iii) examining whether patient neighborhoods implicitly captured finer-grained admission destinations (floor and ICU), even though the model was trained for binary ED disposition. Among admitted patients, 78% were admitted to the floor, while 20% required ICU care.

B. RAPID-TBI Retriever

Upon training, we encode the training and validation patients into multimodal embeddings, which are stored in a FAISS [28] database, a library designed for efficient similarity search across large collections of embeddings. During inference (Fig.1b), a new test patient is encoded into an embedding, and cosine similarity is used to retrieve the top-k similar patients from the FAISS database. These retrieved patients then serve as context for the classifier. For comparison, we also evaluate retrieval performance using baseline k-nearest neighbors search.

C. RAPID-TBI Classifiers

1) Retrieved Patient-Embedding Fused Classifier (RPEC): Using the RAPID-TBI retriever, we first identify the top-k similar patients for a given test case. RPEC uses the test patient's embedding together with the embeddings, class labels, and similarity scores of these retrieved neighbors to improve classification accuracy. This approach models non-linear neighbor relationships while considering class-weighted majority influence. While testing a range of k values, we cap the number of possible retrieved patients at 9, aligning with Miller's Law [35] to ensure cognitive interpretability.

Let $z_p \in \mathbb{R}^{128}$ be the test patient embedding, with $P = \{p_1, p_2, \ldots, p_k\}$ its k-nearest-neighbor embeddings. Each neighbor $p_i \in \mathbb{R}^{128}$ is associated with a binary ED disposition, represented as a one-hot encoded vector $c_i \in \mathbb{R}^2$.

The weight w_i for each neighbor p_i is determined by the cosine similarity with the query embedding, normalized across the k neighbors. The weighted average of neighbor embeddings $p_{\text{agg}} \in \mathbb{R}^{128}$ and the weighted average of the neighbor class vectors $c_{\text{agg}} \in \mathbb{R}^2$ are computed (Eq. 5).

$$p_{\text{agg}} = \frac{\sum_{i=1}^{k} w_i p_i}{\sum_{i=1}^{k} w_i}, \quad c_{\text{agg}} = \frac{\sum_{i=1}^{k} w_i c_i}{\sum_{i=1}^{k} w_i}. \tag{5}$$

We concatenate z_p, p_{agg}, and c_{agg} to form the input to an MLP (g), which predicts the ED Disposition $\hat{y}$ (Eq.6) using binary cross entropy loss.

$$\hat{y} = g\left(z_p \bigotimes p_{\text{agg}} \bigotimes c_{\text{agg}}\right) \tag{6}$$
$$\text{where } [z_p \bigotimes p_{\text{agg}} \bigotimes c_{\text{agg}}] \in \mathbb{R}^{258}$$

2) SLM Classifier: We investigate whether SLMs can serve as prompt-based classifiers to predict ED disposition directly from textual patient summaries through in-context learning. Given the retrieval of top-k similar patients using the RAPID-TBI retriever, we construct a prompt comprising these retrieved patient exemplars followed by the test patient summary (Supplementary Material[1]; Fig. S1), along with a task-specific instruction. Each retrieved patient exemplar includes a textual summary and its corresponding disposition label, serving as few-shot examples to provide the language model with relevant context. The prompt used for the SLM classifier is provided in Supplementary Fig. S2. The SLM outputs a classification

[1] Supplementary Material Available at: `https://github.com/PranavM98/Trust-Your-Neighbors/blob/main/Trust_Your_Neighbors_Supplementary.pdf`

label (e.g., 'Admit to Hospital' or 'Discharge Home') based on the pattern it infers from these patient summaries.

IV. Experiment Design

Supplementary Section-B details descriptions of the training procedures for the encoder, RPEC classifier, and retrieval modules. All results are reported on the held-out test set. Each classification experiment was repeated five times, and we report the mean performance and 95% CI, along with statistical significance relative to RAPID-TBI using a paired t-test. We compare our proposed MMANet encoder against three alternatives: a baseline MMNet (MMANet without attention), the Meta-Transformer [36], and state-of-the-art multimodal architectures for TBI outcome prediction [8], [37], [38]. Supplementary Section-B.1 provides additional implementation details for the baseline comparison experiments. Performance is evaluated using AUROC, F1 score, and Sensitivity. AUROC is our primary metric because it offers a threshold-independent assessment that is robust to class imbalance; Sensitivity and F1 score are taken at the optimal threshold (Youden's index) and provide complementary insight into the models' ability balance recall and precision.

We evaluate two open-source SLMs: Phi-4-mini-instruct (3.8B parameters) [39] and Qwen-2.5-instruct (3B parameters) [40]. Details of the experimental setup are provided in Supplementary Section-B.2. In each prompt, the SLM is explicitly instructed not to provide an explanation but to output only the final classification, whether the patient should be admitted or discharged. As SLMs generate discrete predictions rather than calibrated probabilities, performance is reported using F1 score and Sensitivity instead of AUROC. For each prompting strategy, we vary the contextual examples provided to the model: zero-shot (no patient examples), few-shot-retrieved (patients retrieved by RAPID-TBI retriever), and few-shot-random (randomly selected patient cases).

To evaluate the generalizability of our approach, we conducted two complementary experiments: a cross-institutional test, where models trained on one hospital's data were evaluated on another, and a temporal test, where models trained on earlier years were evaluated on later cohorts (Supplementary Section-B.3 for detailed setup and statistics). Code available: `https://github.com/PranavM98/Trust-Your-Neighbors/tree/main`

V. Results

RAPID-TBI pipeline demonstrated very low latency. On average, utilizing our GPU, patient encoding took 2.0 × 10^{-3} seconds, retrieval of the top 7 similar patients required 1.1 × 10^{-4} seconds per patient, and classification with RPEC averaged 3.5 × 10^{-4} seconds per patient. In contrast, SLM-based predictions with 7 retrieved patients were slower, averaging 0.3 seconds per patient.

TABLE I

Unimodal vs Multimodal Performance. Unimodal uses the generated embeddings, classified using an MLP. (*) indicates p < 0.05. RAPID-TBI results are with 7 retrieved patients. M = Modality where U is Unimodal and MM is Multimodal

Model	**M**	**AUROC**	**F1**	**Sensitivity**
Image [41]	U	0.618 (*) (0.617-0.619)	0.451 (*) (0.436-0.467)	0.538 (*) (0.533-0.542)
Report	U	0.749 (*) (0.748-0.749)	0.547 (*) (0.546-0.547)	0.660 (*) (0.622-0.698)
Tabular	U	0.833 (*) (0.833-0.833)	0.645 (*) (0.645-0.647)	0.816 (*) (0.815-0.817)
Hibi et al. [8]	MM	0.834 (*) (0.833-0.835)	**0.648 (*)** **(0.645-0.652)**	0.819 (*) (0.8096-0.8283)
Pease et al. [38]	MM	0.839 (*) (0.839-0.839)	0.638 (*) (0.636-0.639)	0.850 (0.843-0.857)
Xiong. et al [37]	MM	0.840 (*) (0.840-0.840)	0.645 (*) (0.642-0.650)	0.840 (0.829-0.852)
RAPID-TBI (RPEC)	**MM**	**0.850** **(0.850-0.851)**	0.643 (0.641-0.645)	0.850 (0.836-0.859)
RAPID-TBI (SLM)	**MM**	-	0.629 (0.629-0.629)	**0.871** **(0.871-0.871)**

TABLE II

Few-shot prompting with retrieved similar patients improves classification performance over random examples for both Phi and Qwen.

Prompting Strategy	**Phi**		**Qwen**	
	F1	Sensitivity	F1	Sensitivity
Zero-Shot	0.539	0.602	0.383	0.328
1-Shot	0.555	0.668	0.522	0.483
1-Shot-Random	0.533	0.601	0.482	0.475
3-Shot	0.585	0.789	0.580	0.534
3-Shot-Random	0.529	0.758	0.432	0.387
5-Shot	0.621	0.832	**0.642**	0.665
5-Shot-Random	0.518	0.694	0.485	0.466
7-Shot	**0.629**	**0.871**	0.627	**0.671**
7-Shot-Random	0.501	0.806	0.525	0.582

A. RPEC Classification Results

Table I highlights that RPEC classifier with RAPID-TBI encoder and retriever achieves the highest AUROC (0.850 [0.850–0.851]) among all models, outperforming both unimodal and multimodal baselines. While F1 scores vary across approaches, RAPID-TBI maintains a relatively high F1 score and the highest sensitivity (RPEC: 0.850, SLM: 0.871), striking a strong balance between recall and precision in identifying patients needing admission-minimizing false negatives. For RPEC, retrieving seven patients achieves the highest AUROC and second-highest F1, representing the optimal k for this task (Supplementary Section-F).

Ablation results, provided in Supplementary Table S1, highlight three key findings: (i) RPEC consistently outperforms GraphSAGE and Nearest Neighbors across architectures and encoders; (ii) the MMANet-CTNet configuration for RPEC achieves the highest overall performance; and (iii) using full report text yields better results than binary report labels.

B. SLM Classification Results

Across both SLMs, a consistent pattern emerges: prompting with similar examples improves F1 and sensitivity scores. Compared to zero-shot and few-shot-random baselines, retrieved exemplars consistently enhance model performance, whereas randomly selected examples reduce it (Table II).

Although Qwen achieves the highest F1 score (0.642), it has notably lower sensitivity than Phi, making Phi the more

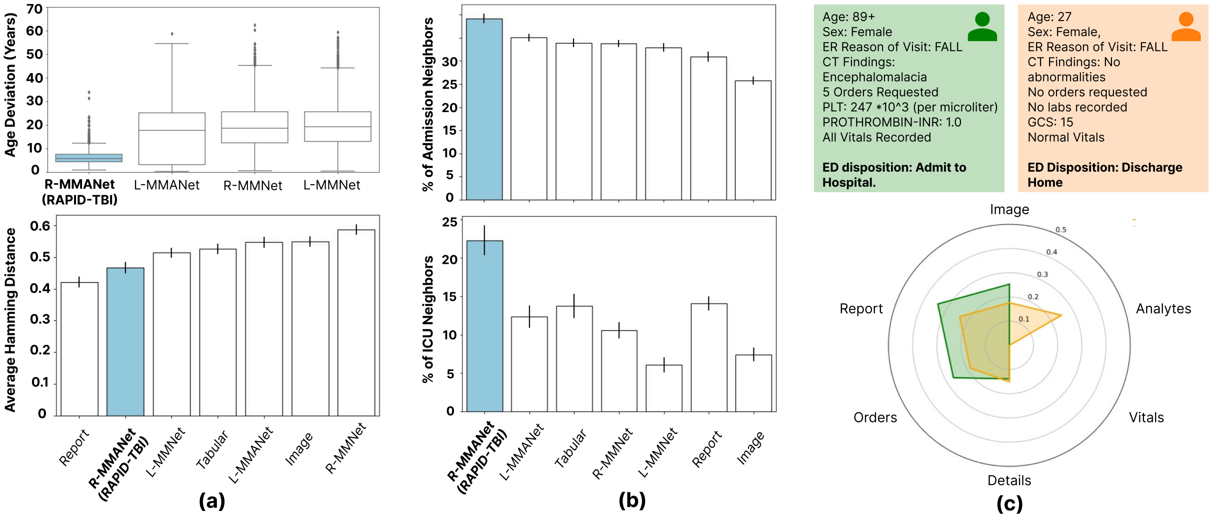

Fig. 3. K=7 for all plots and the error bars represents the standard deviation. (a) Average deviation of age (top) and average hamming distance of x_L (bottom) between test patient and similar neighbors. (b) Percentage of neighbors admitted to the floor (top) or ICU (bottom) when the test patient was also admitted there. (c) Spyder plot of a qualitative example of two patient presentations, illustrating the individualized modality weights assigned by the RAPID-TBI encoder.

reliable model overall. Using seven retrieved similar patients (few-shot prompting) results in the best F1 and sensitivity scores among the prompting strategies. While slightly below the performance of the RPEC classifier, this in-context learning approach comes remarkably close, highlighting the potential of retrieval-augmented prompting for effective prediction.

C. Stratified Results across Patient Subgroups

Supplementary Table S2 illustrates performance of the RPEC classifier across various patient subgroups. Encounters with missing subgroup data were excluded, with final counts shown in the table. The model achieved the highest AUROC for mild TBI (0.820; n = 1316), followed by moderate (0.774) and severe (0.593) cases. Performance across sex was similar, with AUROC values of 0.858 for females and 0.841 for males. Across age groups, AUROC was highest for 36–64 years (0.858), followed by 18–35 years (0.816) and 65+ years (0.802), indicating strongest predictive discrimination in middle-aged patients.

D. Generalizability Evaluation Results

We evaluated generalizability using both the RPEC and SLM classifiers, as detailed in Supplementary Section-E. Despite slight performance degradation under distributional shifts, RPEC remained more robust than GraphSAGE and Nearest Neighbors, achieving AUROCs of 0.727 and 0.764 in cross-institutional and temporal settings, respectively. Although AUROC improved under temporal validation, the F1 score worsened (0.554 vs. 0.586 for cross-institutional), suggesting temporal drift. SLM classifiers showed a similar trend, with stronger cross-institutional generalizability; retrieval at $k = 5$ yielded the best F1 scores, with Qwen (0.627) and Phi (0.622) outperforming their temporal counterparts (0.592 and 0.575). Across both experiments, SLMs outperformed RPEC, demonstrating greater robustness of language models to both temporal and institutional shifts.

E. RAPID-TBI Retrieval Results

To ensure that RAPID-TBI retrieves similar patients, we define “similar patients” as those with comparable ages and CT findings, quantified using the deviation in age and the Hamming distance between their binary CT-finding labels. Although age is provided as an input feature to all multimodal encoders, the MMANet encoder, when supplemented with radiology report information (R-MMANet), retrieves neighbors with a significantly lower mean age deviation compared to other models, representing a statistically significant improvement (Fig. 3a, top).

MMANet encoder demonstrates the added value of leveraging full radiology reports rather than relying solely on extracted labels. Despite not receiving report labels explicitly, it achieves a lower average Hamming distance than label-supervised models like L-MMNet and L-MMANet, suggesting that unstructured text provides richer clinical context for learning meaningful patient representations. (Fig.3a, bottom). Fig.3b further demonstrates its ability to encode latent variables relevant to ED disposition. The top panel shows the percentage of retrieved neighbors also admitted to the floor when the test patient was, and the bottom panel shows the same for ICU admissions. RAPID-TBI encoder outperforms both unimodal and multimodal encoders, retrieving on average 4% more floor-admission neighbors and 10% more ICU-admission neighbors than the next best model.

VI. Discussion

RAPID-TBI aims to enable individualized predictions in two key ways: (i) by learning patient-specific modality attention weights that prioritize the most relevant clinical data for each case, and (ii) by retrieving the most similar patients (neighbors) based on shared clinical features, ensuring that predictions are grounded in comparable past cases.

Building on this foundation, RAPID-TBI outperforms prior state-of-the-art multimodal and graph-based models for TBI prognosis, achieving the highest classification performance for ED disposition. This label is an actual clinical decision documented in the EHR, and while such decisions vary across clinicians and institutions, they reflect the real-world practice patterns that directly govern patient care and resource use [42]. We believe that this variability makes prediction valuable: by learning from aggregate clinician judgments, our model captures the patterns influencing disposition in practice. RAPID-TBI leverages these patterns to provide interpretable retrieval-based reasoning, and our institutional and temporal generalizability evaluations show robustness despite heterogeneity [43]. Interestingly, both classifiers generalize better across hospitals than over time. The greater robustness of SLMs likely reflects their large-scale pretraining on diverse data, whereas RPEC's task-specific training makes it more prone to overfitting and less adaptable to data shifts.

The retrieval-based design offers a practical advantage, as updating the embedding database allows predictions to reflect contemporary standards without retraining the entire model. In addition, retrieval rules or filters can be applied to preferentially select patients from recent time periods, ensuring that case-based evidence remains aligned with current practice. To preserve data fidelity, RAPID-TBI leverages NeuMiss encoders to handle missing values natively, avoiding imputation and reflecting real-world clinical data patterns. Patient-specific attention weights further enhance both performance and interpretability by dynamically focusing on the most informative modalities per case. As illustrated in Fig. 3c, the model places greater emphasis on CT features for a patient with encephalomalacia than for one without any abnormal findings, and at times, identifies missingness as informative. Across the test set, radiology reports received the highest average attention weight (28.7%), followed by orders (22.8%), imaging (21.6%), demographics (15.2%), analytes (7.8%), and vitals (4.0%). Through this dual mechanism, adaptive attention and case-based retrieval, RAPID-TBI delivers personalized, context-aware predictions, laying the foundation for future work to assess its impact on clinical decision-making. Beyond predictive support, retrieved patients enhance model interpretability and trustworthiness. Closely aligned cases provide contextual precedents that substantiate the model's output, whereas markedly dissimilar cases may indicate out-of-distribution inputs or reveal previously unrecognized clinical presentations within the EHR.

Our results also show that, even without fine-tuning, open-source SLMs can approach the performance of trained classifiers like RPEC when provided with relevant in-context examples, specifically, similar patients. This highlights the value of accurate patient retrieval in enabling effective prompt-based classification. Interestingly, while Qwen and Phi show similar F1 scores, Phi achieves higher sensitivity, more often admitting patients who were hospitalized, whereas Qwen tends to miss admissions, potentially reflecting intrinsic differences or biases in their training data during model development.

A. Discrepancy Analysis

We analyzed misclassified test cases to better understand model errors. Misclassified patients were typically older and showed greater age differences from their retrieved neighbors (Supplementary Fig. S5). The retrieved patients' disposition labels were more consistent with the model's predicted label than with the true label, having a mean nDCG of 0.811 (Supplementary Section-G).

B. Clinical Implementation

RAPID-TBI is envisioned as a clinician-assistive tool that integrates seamlessly with existing hospital systems to provide transparent, evidence-based decision support (Supplementary Fig. S8). In practice, such a system, at bed-side, could encode multimodal data in real time, retrieve similar past cases, and generate concise summaries highlighting outcomes, modality contributions, and predicted disposition. With integration through Fast Healthcare Interoperability Resources (FHIR), this vision aims to deliver context-aware, case-based insights at the point of care (Supplementary Section-H).

C. Limitations and Future Work

While RAPID-TBI focuses on clinical data, non-medical factors, such as insurance status, socioeconomic conditions, and hospital capacity, also influence ED disposition decisions, highlighting a key limitation. While we evaluated two forms of generalizability, our data, though drawn from two hospitals, originate from the same health system. Future work should examine RAPID-TBI across multiple, independent health systems to more fully assess its robustness. For the classifiers, SLMs also do not produce calibrated probabilities, preventing AUROC computation and limiting confidence assessment—both critical for clinical use. Future work includes curating a multi-institutional TBI dataset to enable external validation, and clinician user studies to assess impact on decision-making. As retrieval is a major aspect of this system, future work will focus on incorporating fairness measures into the retrieval process [44], to ensure representative patient retrieval that augments care while mitigating potential demographic and practice-related biases in the data. For SLMs, exploring their reasoning when paired with retrieval and evaluating clinical validity through expert review remains a valuable direction.

VII. Conclusion

In this study, we present RAPID-TBI, a multimodal, example-based XAI system for personalized TBI prognosis. Our approach outperforms SOTA multimodal and graph-based models in classification while enabling the retrieval of similar patients. Leveraging a large, curated multimodal TBI dataset, RAPID-TBI integrates measurements from multiple time points during the ED visit and is designed to robustly handle missing data. Modality-specific attention weights further enhance interpretability by highlighting the most relevant data sources for each patient. We also demonstrate the early potential of open sourced SLMs as classifiers that incorporate retrieved patient examples to improve ED disposition prediction. RAPID-TBI thus marks a step toward precision health, where retrieved examples are selected based on multimodal clinical similarity, supporting more informed, personalized predictions by both MLP-based classifiers and language models.

Acknowledgements

We thank Dr. Bradley Kolls for providing clinical insight for this study. This work was supported by the National Institutes of Health (NIH) grant R01NS123275. The study was approved by the Duke Health Institutional Review Board (Protocol Pro00103826).

References

[1] N. W. Siebers and L. A. Steiner, "Anesthesia for traumatic brain injury," *Current Opinion in Anaesthesiology*, vol. 37, no. 5, Oct. 2024.

[2] S. Mehta, L. Devito, E. M. Patsakos *et al.*, "Updated Canadian Clinical Practice Guideline for the Rehabilitation of Adults With Moderate to Severe Traumatic Brain Injury: Mental Health Recommendations," *The Journal of Head Trauma Rehabilitation*, vol. 39, no. 5, Oct. 2024.

[3] R. A. Rutkowski, M. Salwei, H. Barton *et al.*, "Physician Perceptions of Disposition Decision-making for Older Adults in the Emergency Department: A Preliminary Analysis," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting*, vol. 64, no. 1, pp. 648–652, Dec. 2020.

[4] CDC, "Economics of Injury and Violence Prevention," Dec. 2024, https://www.cdc.gov/injury-violence-prevention/economics/index.html.

[5] E. W. Steyerberg, N. Mushkudiani, P. Perel *et al.*, "Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics," *PLoS Medicine*, vol. 5, no. 8, p. e165, Aug. 2008.

[6] D. M. Panczykowski, A. M. Puccio, B. J. Scruggs *et al.*, "Prospective independent validation of IMPACT modeling as a prognostic tool in severe traumatic brain injury," *Journal of Neurotrauma*, vol. 29, no. 1, pp. 47–52, Jan. 2012.

[7] B. Rohaut, C. Calligaris, B. Hermann *et al.*, "Multimodal assessment improves neuroprognosis performance in clinically unresponsive critical-care patients with brain injury," *Nature Medicine*, vol. 30, no. 8, Aug. 2024.

[8] A. Hibi, M. D. Cusimano, A. Bilbily *et al.*, "Development of a Multimodal Machine Learning-Based Prognostication Model for Traumatic Brain Injury Using Clinical Data and Computed Tomography Scans: A CENTER-TBI and CINTER-TBI Study," *Journal of Neurotrauma*, vol. 41, no. 11-12, Jun. 2024.

[9] F. Nasrallah, J. Bellapart, J. Walsham *et al.*, "PREdiction and Diagnosis using Imaging and Clinical biomarkers Trial in Traumatic Brain Injury (PREDICT-TBI) study protocol: an observational, prospective, multicentre cohort study for the prediction of outcome in moderate-to-severe TBI," *BMJ Open*, vol. 13, no. 4, Apr. 2023.

[10] A. Tritt, J. K. Yue, A. R. Ferguson *et al.*, "Data-driven distillation and precision prognosis in traumatic brain injury with interpretable machine learning," *Scientific Reports*, vol. 13, no. 1, Dec. 2023.

[11] M. Amiri, P. M. Fisher, F. Raimondo *et al.*, "Multimodal prediction of residual consciousness in the intensive care unit: the CONNECT-ME study," *Brain: A Journal of Neurology*, vol. 146, no. 1, Jan. 2023.

[12] J. K. Yue, N. Krishnan, J. H. Kanter *et al.*, "Neuroworsening in the Emergency Department Is a Predictor of Traumatic Brain Injury Intervention and Outcome: A TRACK-TBI Pilot Study," *Journal of Clinical Medicine*, vol. 12, no. 5, Mar. 2023.

[13] S. Reddi, S. Thakker-Varia, J. Alder *et al.*, "Status of precision medicine approaches to traumatic brain injury," *Neural Regeneration Research*, vol. 17, no. 10, Feb. 2022.

[14] S. B. Rosenbaum and M. L. Lipton, "Embracing chaos: the scope and importance of clinical and pathological heterogeneity in mTBI," *Brain Imaging and Behavior*, vol. 6, no. 2, Jun. 2012.

[15] R. A. Stocker, "Intensive Care in Traumatic Brain Injury Including Multi-Modal Monitoring and Neuroprotection," *Medical Sciences*, vol. 7, no. 3, Feb. 2019.

[16] S. Yazdani and M. Hoseini Abardeh, "Five decades of research and theorization on clinical reasoning: a critical review," *Advances in Medical Education and Practice*, vol. 10, pp. 703–716, Aug. 2019.

[17] J. Hou and L. L. Wang, "Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences," Feb. 2025, arXiv:2502.20478 [cs].

[18] Z. Wang, Y. Zhu, J. Gao *et al.*, "Retcare: Towards interpretable clinical decision making through LLM-driven medical knowledge retrieving," in *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

[19] Z. Zhang, "RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations," *Optimizations in Applied Machine Learning*, vol. 4, no. 1, Dec. 2024.

[20] M. Fontes, J. D. S. De Almeida, and A. Cunha, "Application of Example-Based Explainable Artificial Intelligence (XAI) for Analysis and Interpretation of Medical Imaging: A Systematic Review," *IEEE Access*, vol. 12, 2024.

[21] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs *et al.*, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, Jul. 2022.

[22] S. G. Anjara, A. Janik, A. Dunford-Stenger *et al.*, "Examining explainable clinical decision support systems with think aloud protocols," *PLOS ONE*, vol. 18, no. 9, Sep. 2023.

[23] A. Defilippo, P. Veltri, P. Lió *et al.*, "Leveraging graph neural networks for supporting automatic triage of patients," *Scientific Reports*, vol. 14, May 2024.

[24] P. Manjunath, B. Lerner, and T. Dunn, "Towards Interactive and Interpretable Image Retrieval-Based Diagnosis: Enhancing Brain Tumor Classification with LLM Explanations and Latent Structure Preservation," in *Artificial Intelligence in Medicine*, 2024.

[25] D. k. Gurmessa and W. Jimma, "Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review," *BMJ Health & Care Informatics*, vol. 31, no. 1, Feb. 2024.

[26] K. Singhal, S. Azizi, T. Tu *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, Aug. 2023.

[27] M. Al-Garadi, T. Mungle, A. Ahmed *et al.*, "Large Language Models in Healthcare," Apr. 2025, arXiv:2503.04748 [cs].

[28] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," Feb. 2017, arXiv:1702.08734 [cs].

[29] M. Kostiuk and B. Burns, "Trauma Assessment," in *StatPearls*, 2025.

[30] R. L. Draelos, D. Dov, M. A. Mazurowski *et al.*, "Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes," *Medical Image Analysis*, vol. 67, Jan. 2021.

[31] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 2019, arXiv:1807.03748 [cs].

[32] E. Alsentzer, J. R. Murphy, W. Boag *et al.*, "Publicly Available Clinical BERT Embeddings," Jun. 2019, arXiv:1904.03323 [cs].

[33] M. Le Morvan, J. Josse, T. Moreau *et al.*, "NeuMiss networks: differentiable programming for supervised learning with missing values." in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5980–5990.

[34] P. Khosla, P. Teterwak, C. Wang *et al.*, "Supervised Contrastive Learning," Mar. 2021, arXiv:2004.11362 [cs].

[35] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[36] Y. Zhang, K. Gong, K. Zhang *et al.*, "Meta-Transformer: A Unified Framework for Multimodal Learning," Jul. 2023, arXiv:2307.10802 [cs].

[37] Z. Xiong, K. Zhao, L. Ji *et al.*, "Multi-modality 3D CNN Transformer for Assisting Clinical Decision in Intracerebral Hemorrhage," in *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2024, pp. 522–531.

[38] M. Pease, D. Arefan, J. Barber *et al.*, "Outcome Prediction in Patients with Severe Traumatic Brain Injury Using Deep Learning from Head CT Scans," *Radiology*, vol. 304, no. 2, pp. 385–394, Aug. 2022.

[39] M. Abdin, J. Aneja, H. Behl *et al.*, "Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs," Mar. 2025, arXiv:2503.01743 [cs].

[40] Qwen, A. Yang, B. Yang *et al.*, "Qwen2.5 Technical Report," Jan. 2025, arXiv:2412.15115 [cs].

[41] I. E. Hamamci, S. Er, F. Almas *et al.*, "Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography," Oct. 2024, arXiv:2403.17834 [cs].

[42] I. G. Stiell, G. A. Wells, K. Vandemheen *et al.*, "The Canadian CT Head Rule for patients with minor head injury," *The Lancet*, vol. 357, no. 9266, pp. 1391–1396, May 2001.

[43] N. Carney, A. M. Totten, C. O'Reilly *et al.*, "Guidelines for the Management of Severe Traumatic Brain Injury, Fourth Edition," *Neurosurgery*, vol. 80, no. 1, pp. 6–15, Jan. 2017.

[44] F. Chen and H. Fang, "FAIR-QR: Enhancing Fairness-aware Information Retrieval through Query Refinement," Mar. 2025, arXiv:2503.21092 [cs].