

An Explainable Hybrid GraphRAG Framework for Diabetes Risk Stratification

Herman Kandolo*, Rabbi Kimbangi Menakuntima*, Wenchel Ridore*, Heritier Hamba Mawandu*
Frederic Habonimana*, Alain M. Kuyunsa†, Grevi Nkwimi‡

*International School, Vietnam National University (IS-VNU), Hanoi, Vietnam

†Université de Kinshasa, Kinshasa, Democratic Republic of the Congo

‡Dept. of Applied Artificial Intelligence, Hanyang University, ERICA Campus, Ansan-si, Republic of Korea

Corresponding author: planeteherman@gmail.com

Abstract—Clinicians and public health practitioners increasingly need decision-support tools that are not only accurate but also transparent enough to be trusted at the point of care. This paper presents an explainable clinical decision support framework for diabetes risk stratification, built on a hybrid GraphRAG architecture that shifts statistical reasoning away from the large language model (LLM) and into a deterministic knowledge graph, so that the LLM’s role is strictly limited to generating a clinician-readable explanation of a pre-computed risk profile rather than performing the computation itself. Patients are stratified into four clinically interpretable risk profiles (low risk, vascular risk, severe cardiovascular risk, high diabetes prevalence) via unsupervised clustering, routed in real time through a lightweight classifier, and matched against a Neo4j knowledge graph that returns fully auditable statistical evidence for each recommendation – addressing a core barrier to clinical adoption of AI-based e-health tools: the inability of clinicians to verify why a system reached a given conclusion. Evaluated on 70,692 records from the BRFSS 2015 dataset, the framework achieves 99.2% routing accuracy (binary high-risk vs. low-risk discrimination AUC of 0.9997) and, compared against an implemented vector-based RAG baseline, transmits substantially less context to the LLM while eliminating numerical hallucinations entirely (0% versus 19.2% for the baseline), a property potentially relevant to safety-oriented e-health deployment. Unlike vector-based RAG, where semantic similarity offers no interpretable traceability, every decision step in the proposed pipeline produces an artifact independently verifiable by a clinician, shifting explainability from a post-hoc justification to a structural, by-design property – directly addressing the trust and transparency requirements of pervasive e-health decision-support systems.

Keywords—Clinical decision support, explainable AI, knowledge graphs, GraphRAG, e-health, diabetes risk stratification, large language models

I. INTRODUCTION

Type 2 diabetes does not progress quietly. It advances in plain sight in global health statistics: an estimated 537 million adults were living with diabetes worldwide in 2021, a figure projected to reach 783 million by 2045 [1], and the complications it causes – cardiovascular disease, stroke, renal failure, vision loss – represent a growing burden on already strained healthcare systems. What has changed recently is the ability of digital tools to intervene earlier, to identify at-risk profiles, and to guide clinicians toward more timely care.

A major barrier to the clinical adoption of AI-based decision support, however, is not accuracy alone but trust [2]: clinicians and patients are reasonably reluctant to act on

a recommendation they cannot verify, and empirical studies of clinician needs consistently identify interpretability as a precondition for adoption rather than a desirable add-on [3]. This is particularly acute in pervasive e-health settings [4], where decisions may be generated far from direct clinical supervision – at home, in a smart clinic, or via a remote monitoring platform – and must therefore be independently auditable rather than taken on faith. The recent emergence of large language models (LLMs) has generated legitimate enthusiasm as a natural-language interface for such systems, capable of comprehending and synthesizing complex information at a scale inaccessible to traditional approaches. Yet LLMs suffer from a fundamental limitation as a computational backbone for clinical decision support: they retain no memory of clinical data between queries, and when tasked with analyzing thousands of patient records, they must receive the entirety of that data at every request – a constraint that translates into high inference costs, degraded response times, and, more critically for patient safety, a non-negligible risk of producing incorrect statistical assertions (hallucinations) [5] that a clinician may not be equipped to catch.

Retrieval-Augmented Generation (RAG) architectures constituted a first response to this problem [6]. Rather than injecting the entirety of available data into the model’s context, these systems dynamically retrieve the most relevant documents for each query through vector-based semantic similarity search. Results are encouraging, but do not resolve everything: in a medical context, retrieved documents remain voluminous, the task of statistical interpretation still rests on the language model, and – critically for clinical trust – a similarity score offers no clinically meaningful explanation of why a particular recommendation was produced.

Our proposal rests on a simple but structuring intuition: if the analytical workload can be shifted to a structured, deterministic data layer, the language model no longer needs to reinvent statistical reasoning at every query, and every step of that reasoning becomes independently inspectable by a clinician. The LLM becomes what it does best – an interpreter, translating verified numbers into clinical narrative – while the underlying risk stratification remains fully auditable, a prerequisite for deployment in pervasive, loosely supervised e-health environments.

The main contributions of this work are:

- An explainable clinical decision support framework

in which every risk stratification decision produces an artifact independently auditable by a clinician, directly addressing a core adoption barrier for AI-based pervasive e-health tools;

- A hybrid GraphRAG architecture integrating unsupervised learning, supervised classification, a deterministic knowledge graph, and LLM-based generation, in which the LLM is strictly confined to explanation rather than computation;
- Construction and clinical characterization of four interpretable diabetes risk profiles from the BRFSS 2015 dataset;
- Experimental evidence, against an implemented vector-based RAG baseline, of a substantially reduced LLM context cost and a 0% numerical hallucination rate versus 19.2% for the baseline – a property relevant to safety-oriented point-of-care use;
- A structural, step-by-step characterization of explainability directly relevant to trustworthy pervasive e-health decision-support systems.

II. RELATED WORK

A. Diabetes Prediction and Patient Clustering

Early diabetes prediction is an active research area, commonly relying on decision trees, random forests, SVMs, and K-NN [7] to identify high-risk individuals. Most such work focuses on predicting a single target class without providing a structured representation of the underlying medical knowledge. Unsupervised clustering, particularly K-Means [8], enables the discovery of patient sub-populations sharing similar risk profiles, though such clusters typically remain descriptive rather than integrated into a real-time decision system.

B. RAG and GraphRAG Architectures

The emergence of LLMs has driven the development of RAG architectures [6], [9], which combine a document retrieval engine with a generative model. This approach improves response relevance but presents critical limitations in medical contexts: high token consumption, retrieval latency, and hallucination risk when interpreting complex statistical data. To address these limitations, several works have proposed GraphRAG architectures integrating knowledge graphs [10], which explicitly represent relationships between medical entities [11], building on a longer tradition of learning structured health knowledge graphs directly from clinical records [12] to improve response consistency while reducing hallucinations. Most existing approaches, however, still rely on hybrid mechanisms combining vectors and graphs, retaining a dependency on voluminous text processing.

C. Positioning

Unlike prior work, our approach coherently integrates unsupervised clustering, supervised classification, a deterministic knowledge graph, and LLM generation into a single explainable decision support pipeline rather than a purely retrieval-oriented one. By shifting analytical intelligence to the graph rather than the language model, every risk stratification step

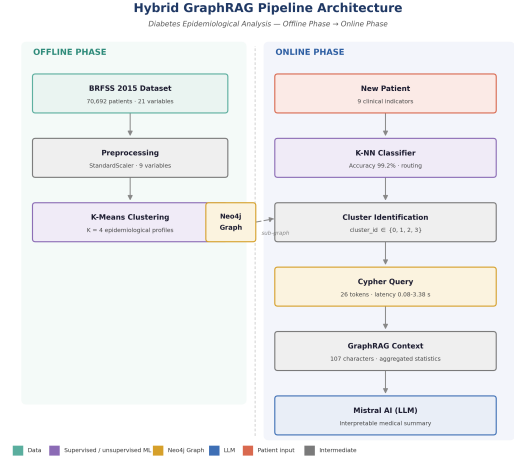


Fig. 1. Overall pipeline of the hybrid GraphRAG architecture. The offline phase (left) builds epidemiological knowledge: BRFSS 2015 dataset → preprocessing → K-Means clustering → storage in the Neo4j graph. The online phase (right) processes each new patient: K-NN classification → cluster identification → Cypher query → compact GraphRAG context → medical summary generation.

remains independently auditable by a clinician, while reducing contextual volume transmitted to the LLM and improving result explainability – validated experimentally rather than estimated.

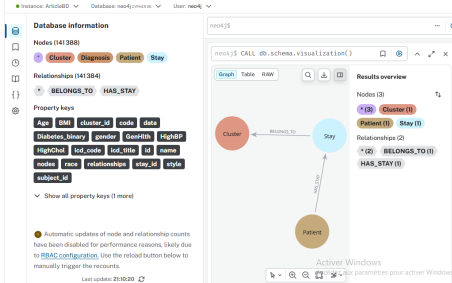
III. PROPOSED METHODOLOGY

A. Architecture Overview

The proposed architecture rests on two complementary phases: an offline phase that constructs epidemiological knowledge, and an online phase enabling instant evaluation of a new patient. The offline phase starts from the BRFSS 2015 dataset, applies preprocessing, then K-Means clustering to identify four epidemiological profiles, loaded into a Neo4j graph. The online phase receives a new patient’s profile, classifies it via K-NN, queries the corresponding sub-graph via a Cypher query, and submits the resulting ultra-compact context to the LLM for medical summary generation (Fig. 1).

B. Dataset

Experiments were conducted on the “Diabetes Binary Health Indicators BRFSS 2015” dataset, from the CDC’s Behavioral Risk Factor Surveillance System [13], containing 70,692 individuals, 21 medical and behavioral variables, and a binary diabetes target. This dataset was selected for its public availability, sufficient size for pipeline evaluation, and frequent use as a benchmark in diabetes prediction literature. Importantly, this variant is a deliberately balanced 50:50 derivative of the original BRFSS survey (oversampling diabetic respondents to match non-diabetic ones), not a population-representative sample. Consequently, the diabetes prevalence figures reported per cluster in Section IV-A (20.8% to 74.4%) reflect relative differences *within this balanced cohort* and must not be interpreted as, or transported to, real-world epidemiological prevalence without appropriate reweighting against the full, unbalanced BRFSS population. Its self-reported nature is a further recognized limitation (Section V-D). Data were normalized using a StandardScaler prior to analysis.



C. Epidemiological Profile Construction (K-Means)

This feature set mixes six binary indicators, one continuous variable (BMI), and two ordinal variables, for which Euclidean distance on standardized values is a common but imperfect simplification – validated against mixed-type alternatives (e.g., Gower distance, K-prototypes) is left for future work and noted as a limitation (Section V-D).

D. Knowledge Graph Modeling (Neo4j)

E. Real-Time Classification (K-NN)

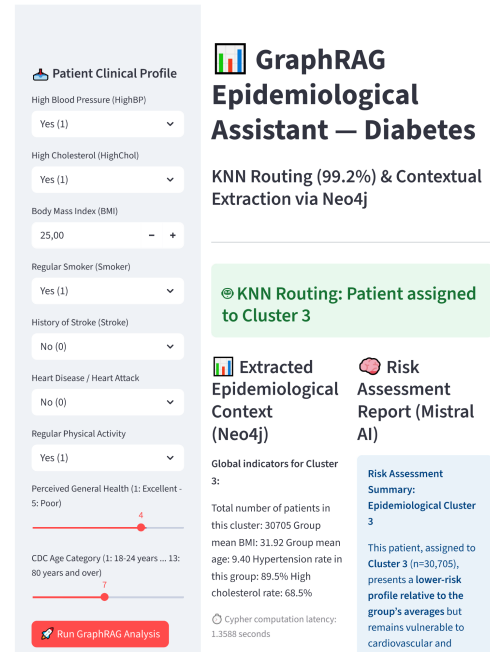


Fig. 4. Epidemiological assistant interface. The left panel collects the patient’s clinical indicators; the right panel displays the cluster identified by K-NN, the context extracted from Neo4j (26 tokens), and the generated medical summary, with a numerical hallucination rate of 0.0%.

profile. It is important to clarify this component’s scope: K-NN is not a classifier intended to generalize to external populations, but an internal routing mechanism trained to replicate the partition defined by K-Means on the same dataset. Its 99.2% performance reflects pipeline-internal consistency rather than external validation.

F. Deterministic Extraction and Generation

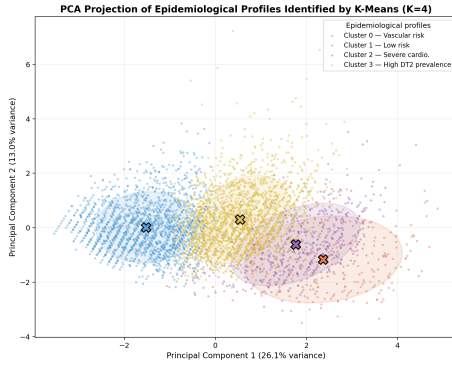


Fig. 5. 2D PCA projection of the four epidemiological profiles identified by K-Means. The first two principal components explain 26.1% and 13.0% of total variance (39.1% cumulative). Cluster 1 (low risk) is clearly separated along the horizontal axis from Clusters 0, 2 and 3 (elevated cardiometabolic risk profiles), which partially overlap.

TABLE I. INTRINSIC AND EXTRINSIC CLUSTERING METRICS

Metric	Value
Silhouette coefficient	0.1667
Davies-Bouldin index	13.0366
Calinski-Harabasz index	8,828.48
Purity	0.7250
Adjusted Rand Index (ARI)	0.1472

IV. EXPERIMENTAL RESULTS

A. Clustering Evaluation

Table I presents intrinsic clustering metrics. While the silhouette coefficient indicates moderate separation, the high Calinski-Harabasz index confirms an exploitable partitioning structure. The purity of 72.5% indicates a non-uniform distribution of diabetes status across the resulting clusters, although the low ARI (0.1472) and silhouette coefficient (0.1667) indicate that this association should not be interpreted as evidence of clinically validated phenotypes (Fig. 5).

B. K-NN Routing Performance

Table II reports per-cluster K-NN performance (80/20 test split, $k = 21$, $n = 14,139$). Cluster 0 (vascular risk, 879 test individuals, 6.2% of the cohort) reaches perfect precision and recall, reflecting its 100% stroke prevalence. This does not imply stroke predicts membership by design; rather, stroke history is comparatively rare among the nine clustering variables, and the small subset reporting it consistently co-occurs with the hypertension/cholesterol/BMI combination defining this cluster's centroid – a descriptive artifact of unsupervised partitioning, not a validated clinical marker. Clusters 1 and 3, larger and more diffuse, still achieve F1 scores above 0.98.

C. Binary Clinical Risk Metrics

To situate our results against standard clinical decision support evaluation practice, we additionally report binary discrimination metrics obtained by collapsing the four clusters into a clinically-justified High Risk / Low Risk label (Cluster 1, low risk, versus Clusters 0, 2 and 3, all presenting elevated cardiometabolic comorbidity and diabetes prevalence, as High Risk), derived directly from the same K-NN model's class probabilities without any retraining. Table III reports these

TABLE II. PER-CLUSTER K-NN PERFORMANCE

Cluster	Profile	Precision	Recall	F1
C0	Vascular risk	1.0000	1.0000	1.0000
C1	Low risk	0.9897	0.9895	0.9896
C2	Severe cardio.	0.9958	1.0000	0.9979
C3	High DT2 prevalence	0.9919	0.9909	0.9914
W. avg	Acc.: 99.20%	0.9920	0.9920	0.9920

TABLE III. BINARY CLINICAL RISK DISCRIMINATION METRICS

Metric	Value
AUC-ROC	0.9997
Sensitivity (high risk)	0.9936
Specificity	0.9895
Positive Predictive Value	0.9934
Negative Predictive Value	0.9897
Accuracy	0.9920

metrics on the same held-out test set ($n = 14,139$); Fig. 6 shows the corresponding ROC curve.

These figures require an important, more explicit caveat than a routine limitation: the reported accuracy, AUC, sensitivity, and specificity are internal consistency measures, not evidence of diagnostic or prognostic performance. The binary high-risk/low-risk label used to compute them is itself derived by collapsing the same K-Means clusters that the K-NN classifier was trained to reproduce (Section V-D); the evaluation is therefore circular by construction, and the near-ceiling AUC reflects the strong separability of that internally generated label structure rather than an independent clinical ground truth such as confirmed future diabetes onset, diagnosed complications, or outcomes adjudicated independently of the clustering pipeline. These metrics should not be presented, cited, or interpreted as clinical discrimination performance. They are reported here only to enable comparison with the reporting conventions of the clinical decision support literature, which typically expects such metrics – not as evidence that the system predicts future diabetes risk.

D. Experimental Comparison Against Vector-Based RAG

Unlike a theoretical estimate, we implemented an actual vector-based RAG system on the same dataset to produce a direct experimental comparison. Each patient is converted into a textual document describing nine clinical variables, encoded with *all-MiniLM-L6-v2*, and indexed in a ChromaDB vector store (stratified sample of 5,000 patients). For each new patient, the 20 most similar documents are retrieved via cosine similarity and concatenated into raw context, as a conventional RAG system would.

Table IV compares both architectures on four representative patient profiles, each latency measurement averaged over ten repetitions. Context reduction averages 97.31% (from ~ 998 to 26 tokens), measured experimentally rather than theoretically estimated. Latency, however, is more nuanced than initially expected: vector-based RAG, querying a locally-cached ChromaDB index, shows a mean latency of 0.1012 s, while the Cypher query underlying GraphRAG shows a mean latency of 0.9920 s, including K-NN classification (~ 25 ms). This latency varies substantially by cluster queried – from 0.08 s to 3.38 s – likely related to Neo4j caching behavior rather than cluster size alone. The robust and consistent advantage of the proposed architecture is therefore not retrieval speed,

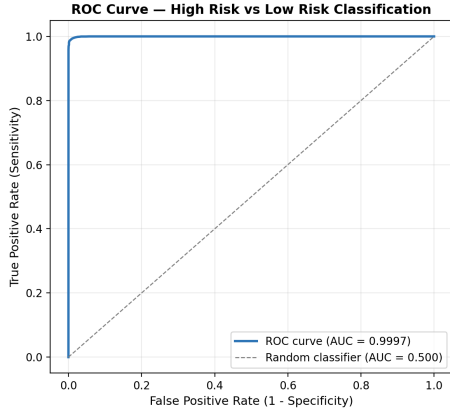


Fig. 6. ROC curve for binary High Risk vs Low Risk classification, derived from K-NN class probabilities (AUC = 0.9997).

TABLE IV. EXPERIMENTAL COMPARISON: VECTOR-BASED RAG VS. GRAPH RAG

Metric	Vector RAG	GraphRAG
Tokens (mean)	998	26
Context reduction	—	97.31%
Latency (mean)	0.1012 s	0.9920 s
Latency (min-max)	0.047–0.189 s	0.11–3.41 s
Knowledge source	Embeddings (top-20)	Neo4j graph
Hallucination rate	19.2% (measured)	0% (measured)
Explainability	Low	High

but the massive and constant reduction of contextual volume transmitted to the language model.

E. Numerical Hallucination Evaluation

For GraphRAG, out of 120 numerical values generated from Neo4j-extracted statistics, no hallucination was detected (0% rate), since the transmitted context contains exactly the statistics the model must interpret. For vector-based RAG, where the LLM must itself aggregate percentages from 20 raw retrieved documents, an initial automated keyword-matching detection attempt proved unreliable in the presence of two recurring linguistic phenomena: negations affecting the percentage itself (e.g., “absence of hypertension: 100%” meaning 0% of patients are hypertensive), and clauses grouping multiple statistics whose associated keywords overlap. Manual validation was therefore conducted on 41 percentages generated across four scenarios: 15 could not be matched to any measurable ground-truth statistic (e.g., obesity rates derived from BMI thresholds, not directly computed in our reference data) and were excluded; among the 26 remaining semantically valid comparisons, 5 significant discrepancies (exceeding 5 percentage points) were identified, yielding a 19.2% numerical hallucination rate. This result, while based on a limited sample, empirically confirms the theoretical risk associated with conventional RAG architectures and contrasts sharply with the zero rate obtained by construction with GraphRAG. Both protocols measure fidelity in repeating or computing figures from provided context, not the model’s capacity to invent values absent any context; qualitative interpretation errors are not covered and constitute a direction for future work.

V. DISCUSSION

A. Clustering Relevance and Routing Reliability

Despite a moderate silhouette coefficient, detailed cluster analysis reveals meaningful functional separation across the four profiles. Cluster 1 (low risk, 7.3% hypertension, 20.8% diabetes) is clearly distinct from Clusters 0, 2, and 3, which partially overlap and concentrate populations with elevated cardiometabolic vulnerability (diabetes prevalence between 65.8% and 74.4%). The K-Means/K-NN pairing forms an efficient internal routing mechanism; its 99.2% performance reflects pipeline consistency rather than external generalization – external validation on an independent dataset remains an important direction for future work.

B. Graph Contribution and Observed Latency Limits

Replacing document retrieval with deterministic graph querying is the central architectural contribution of this work: our approach submits only 107 characters (26 tokens) of pre-computed statistics, against ~4,000 characters in the implemented vector-based RAG on the same dataset. We present Neo4j here as a structured, auditable, and extensible knowledge layer rather than as evidence that graph-native traversal is intrinsically superior to a relational store for the present workload (see Limitations, Section V-D); the measured benefit in this study stems from pre-aggregation and deterministic retrieval, which a well-indexed relational schema could in principle also provide. The model is no longer a calculator but an interpreter, explaining the zero hallucination rate observed for GraphRAG. This contrasts with the dominant paradigm in clinical explainable AI, where post-hoc attribution methods such as SHAP [15] are applied after a black-box model has already produced its output; in the proposed pipeline, the decision is explainable by construction, as each intermediate artifact – cluster assignment, Cypher query, aggregated statistics – is inspectable independently of any dedicated explanation method. Cypher query latency, however, only partially confirms the initial speed promise, varying from 0.08 s to 3.38 s across clusters with substantial variability (std. dev. up to 6.90 s) that persists even after indexing on *cluster_id*. This variability appears more closely linked to Neo4j’s internal caching mechanisms than to cluster size alone – Cluster 1 (27,223 patients) shows stable latency despite its size, while Cluster 3 (30,705 patients) shows the most unstable latency. Vector-based RAG, with its fully memory-resident ChromaDB index, shows consistently sub-second latency (0.05–0.19 s). The robust and constant advantage of the proposed architecture thus lies in contextual volume reduction, not retrieval speed, which would require further Neo4j configuration tuning to become competitive.

C. Computational Cost

The offline K-Means clustering (70,692 records) runs once, not per query. Per-patient online cost comprises K-NN classification (~25 ms), one Cypher query (0.08–3.38 s), and one LLM call (26-token input, *max_tokens*=1000) – the latter dominating end-to-end latency via external API overhead. The 97.31% context reduction versus vector-based RAG proportionally lowers marginal LLM cost, though wall-clock API cost and comparison against a non-LLM templated baseline were not measured; a full cost analysis is left for future work.

D. Limitations

Several limitations merit attention. BRFSS 2015 relies on self-reported data and a limited variable set. Summaries remain descriptive, not diagnostic; hallucination evaluation covers only explicit numerical values. K-NN performance (99.2%) reflects pipeline-internal consistency, not external validation, and its advantage over direct nearest-centroid assignment was not benchmarked experimentally in this study. Clustering shows only moderate intrinsic quality (silhouette 0.1667, Davies-Bouldin 13.04), with visible PCA overlap among elevated-risk clusters (Fig. 5); labels such as “vascular risk” aid readability, not clinician-validated phenotypes. Euclidean distance on standardized mixed-type variables (binary, ordinal, continuous) is a common but imperfect simplification, not validated here against mixed-type alternatives such as Gower distance or K-prototypes; nor was cluster stability across initializations formally assessed. The RAG comparison is not fully task-equivalent, since vector-based RAG aggregates raw records within the LLM call while GraphRAG receives pre-aggregated statistics – part of the advantage stems from *where* aggregation occurs, not retrieval alone; an isolating ablation is future work. The graph currently stores precomputed memberships via direct lookups, without demonstrating multi-hop traversal beyond an indexed table; its case rests on schema flexibility for future extension (e.g., patient-event-biomarker-trial graphs in oncology, Section VI) rather than graph-native reasoning shown here. The RAG baseline uses a 5,000-patient stratified sample (of 70,692) and $n = 10$ repetitions per measurement; larger-scale evaluation and resolving cluster-dependent latency variability remain open. Finally, no clinician assessment, fairness analysis, calibration study, or workflow evaluation was performed; claims of clinical trust and point-of-care suitability are hypotheses for future evaluation, not demonstrated outcomes.

VI. CONCLUSION AND FUTURE WORK

This study proposed an explainable clinical decision support framework for diabetes risk stratification, built on a hybrid GraphRAG architecture combining machine learning, knowledge graphs, and large language models. By confining the LLM to explanation rather than computation, every stratification decision remains independently auditable by a clinician – directly addressing a core barrier to the clinical adoption of AI-based e-health tools. Results obtained on the BRFSS 2015 dataset – 99.2% routing accuracy, substantially reduced experimental context cost against an implemented vector-based RAG baseline, and a numerical hallucination rate of 0% versus 19.2% for that baseline – demonstrate that a substantial portion of statistical reasoning can be shifted to the data layer, reducing the observed risk of numerical inconsistency and improving explainability. Retrieval latency, however, remains to be optimized for certain large clusters, an identified direction for technical work.

Several extensions are envisioned: replacing K-Means with more advanced methods (DBSCAN, HDBSCAN, Gaussian Mixture Models), integrating Graph Neural Networks, and validating on real hospital datasets such as MIMIC-IV or eICU – directly relevant to pervasive e-health deployments where data heterogeneity and real-time constraints are more pronounced than in a static benchmark dataset. Clinical expert

evaluation constitutes an essential next step to assess the clinical relevance and decision utility of generated summaries in real-world pervasive health environments.

REFERENCES

- [1] International Diabetes Federation, “Idf diabetes atlas,” International Diabetes Federation, Brussels, Belgium, Tech. Rep., 2021.
- [2] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, “An overview of clinical decision support systems: Benefits, risks, and strategies for success,” *npj Digital Medicine*, vol. 3, no. 1, p. 17, 2020.
- [3] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, “What clinicians want: Contextualizing explainable machine learning for clinical end use,” in *Proceedings of the 4th Machine Learning for Healthcare Conference (MLHC)*, vol. 106, 2019, pp. 359–380.
- [4] U. Varshney, “Pervasive healthcare and wireless health monitoring,” *Mobile Networks and Applications*, vol. 12, no. 2–3, pp. 113–127, 2007.
- [5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [7] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [8] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
- [9] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [10] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, “From local to global: A GraphRAG approach to query-focused summarization,” *arXiv preprint arXiv:2404.16130*, 2024.
- [11] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*, 2nd ed. O’Reilly Media, 2015.
- [12] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, “Learning a health knowledge graph from electronic medical records,” *Scientific Reports*, vol. 7, no. 1, p. 5994, 2017.
- [13] K. Stanton, “Diabetes health indicators dataset (brfss 2015),” Kaggle, 2021.
- [14] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, and A. Taylor, “Cypher: An evolving query language for property graphs,” in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD)*, 2018, pp. 1433–1445.
- [15] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.