Secure Curation and Reuse of Digital Scientific Data Using Retrieval-Augmented Generation Systems

Harshetha Murthy Keshav Murthy[1], Alexander Iliev Iliev[1, 2][0000-0002-4220-3637]

[1] SRH Hochschule Berlin, Sonnenallee 221, Berlin, Germany
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria, 8 Acad. Georgi Bonchev Str., Bulgaria, 1113, Sofia
harshethabm@gmail.com, ailiev@berkeley.edu

Abstract. This study explores secure and intelligent curation of digital scientific data using Retrieval-Augmented Generation (RAG) systems. Focusing on privacy-preserving approaches and reuse of structured healthcare datasets, we propose models for fraud detection, enhancing semantic interpretation, personalization, and secure access to digital knowledge assets in critical sectors.

Keywords: Retrieval-Augmented Generation (RAG), Intelligent Curation, Digital Scientific Heritage, Privacy-Preserving AI, Creative Reuse of Data.

1 Introduction

The digitalization of culture and science entails benefits and risks for the management of assets in the form of data. Among the sophisticated forms of AI, Retrieval-Augmented Generation (RAG) have the potential to revolutionize information by integrating retrieval and generative models and provide up-to-date, contextually superior, and semantically enchanced output. These capabilities are extremely pertinent in areas that would call for customization and timeliness. However, with increasing adoption in sensitive arears like healthcare, concerns related to privacy, abuse, and malicious attacks have grown. The major issues related to privacy, abuse, and susceptibility to malicious attacks arise (Nazary et al., 2025; Jiang et al., 2024). In this paper, we explore how to safely repurpose structured healthcare datasets based on the RAG architecture. It also acknowledges issues of semantics regulation, smart search information, and the adoption of proper ethical approaches to artificial intelligence. On that note, this work enriches the conference's theme by applying trustworthy artificial intelligence into heritage and data-intensive domains to facilitate sustainable and secure curation of digital scientific assets.

Digital Presentation and Preservation of Cultural and Scientific Heritage, Vol. 15, 2025.
Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. ISSN: 1314-4006, eISSN: 2535-0366

2 Exposition of the Investigation

2.1 Conceptual Framework

Retrieval-Augmented Generation or RAG creates the concept of using search and generative AI but makes the output much more factually and contextually anchored than otherwise. As contrasted with previous language models that are trained solely on predetermined knowledge bases, RAG models can ask for information from outside sources which make it possible to add relevancy and semantic depth in real-time as confirmed by Asai et al., 2023). This is especially so in application areas of the likes of the health sector where accuracy in operations and the data being processed is very crucial. However, when RAG is applied to sensitive fields, the privacy and security factors arise, and data leakage, misuse, and unauthorized access become possible (Chen et al., 2025, Koga et al., 2024). To overcome these challenges, the given framework includes the steps such as Homomorphic encryption and Federated learning in the system. These methods ensure the privacy of the users' data and the decentralization of their information when, at the same time, updates can be made for the semantic reuse of the models (Aly & Smart, 2019; Geren et al., 2024). In this way, this combination facilitates the ethical, secure and scalable implementation of RAG in high-risk environments.

2.2 Methodology and Data Sources

The data collection for the study is public and structured in the context of healthcare fraud analysis. It includes the patient and claims data, inpatient and out-patient, fraud factors that relate to such practices of providers. Before data analysis, some preprocessing that was done were data cleaning, missing records handling by using different algorithms, data encoding for categorical data and record joinery on different sources. This led to the use of feature engineering to get other related features like the number of days the patient stayed in the hospital, the period of the claim, and the multiple roles of the physician. These are to identify instances of fraud such as distortion of facts or inconsistencies in the submissions by the providers, as stated by Lavingia and Mehta (2022). The last dataset was pre-processed so that the number of samples in each class is equal to 50,000, as well as normalized for the models and for the interpretation.

2.3 Machine Learning Models

Three models of machine learning were trained and tested for the purpose of fraud detection:

- Logistic Regression also worked as another classifier; it was simple and easy to understand.
- Decision Trees or Random Forests made the decision-making process more transparent and this is important especially when it comes to audit.

- Choice of XGBoost as the key model was due to its high reliability and effectiveness in classifications.

Models were evaluated using standard parameters for successes and failures; accuracy, precise, recall, and F1 score. Chi-Square test was used in performing feature selection while cross-validation was used to split the dataset to check the generality and to reduce model overfitting.

3 Results

Fraud detection in the healthcare sector is a very important task since it helps in preventing fraudulent claims and losses to the health care systems. The identification of the fraudulent providers is not an easy task, which is why it needs to use the complex methods of analysis. This chapter outlines a detailed procedure on how to analyze the claims data to identify fraud using machine learning. The paper uses the claim data elements and builds a fraud detection model to detect the provider's abnormal activity, which may involve fraud. Therefore, in order to meet the above objective, the study adopts the following steps: data cleaning, data preparation, EDA, and model building. Data preprocessing comprises of combining more than one dataset, dealing with missing data, as well as categorizing nominal data for the machine learning models.

Fraud detection in healthcare is a significant challenge where classification of the providers that could be involved in fraud is important. This work employs the use of machine learning techniques in order to classify the structured claim data for identifying fraudulent providers. In order to achieve this goal, we used data pre-processing, feature engineering, exploratory data analysis and model training.

3.1 Feature Engineering

The code performs feature engineering by creating two new features: hospital stay duration and physician role consistency.

The first part is the number of days the patient spent in the hospital, arriving at Hperiod by converting AdmissionDt and DischargeDt into datetime format. Another characteristic, claim amount (claim), is obtained from the InscClaimAmtReimbursed column, and the claim period (period) is obtained in the same manner using ClaimStartDt and ClaimEndDt. The second part consists of a function physician_same that identifies AttendingPhysician, OperatingPhysician, and OtherPhysician. It provides different values depending on whether the physician is listed in more than one role. The feature phy_same can be used in detecting possible fraud where a provider submits multiple entries of the same physician to make more claims.

This code is intended to use label Encoding for changing the PotentialFraud variable into the numerical form. As shown above, it replaces "Yes" with 1 and "No" with 0, which is ideal for the machine learning models. This encoding is crucial for the algorithms that work with the number input, which can facilitate the identification of frauds based on the provider's data.

3.2 Feature Scaling

This code carries out the Standard Scaling with StandardScaler() from sklearn.preprocessing package to normalize numerical values. The claim column is scaled to have zero mean and unit variance, so all the numerical variables are on the same level of measurement. It is used in this process to scale down features with large values so that they do not influence the training process for the model. The transformed values are assigned to a new variable in the form of DataFrame, replacing the original column. Standardizing features is important in improving the models, especially those which are affected by the magnitude of the features.

3.3 Exploratory Data Analysis

Fig. 1. Count Plot of distribution of fraud case.

The count plot is to show the distribution of the fraud cases with the help of seaborn's countplot (). It charts the number of the fraudulent (1) and non-fraudulent (0) providers obtained from PotentialFraud. The bar chart shows class imbalance, with more non-fraudulent cases than fraudulent ones. That is where this imbalance comes in handy when selecting methods such as oversampling or weighted models to increase the chances of detecting fraud.

Scatter Plot of Claims vs. Reimbursement Amount

This scatter plot shows the insurance claim amounts (claim) against the IPAnnualReimbursementAmt and also differentiate between fraudulent (PotentialFraud=1) and non-fraudulent (PotentialFraud=0) cases. The x-axis denotes claims and the y-axis denotes the annual reimbursement. It shows that fraudulent claims (marked as orange points) involve higher reimbursement amounts than the probable normal range and therefore can be considered as anomalous. This makes the identification of various patterns of the claim behavior, such as in the cases of high-claim and high-reimbursement to be easier and thus making it easier to detect fraud.

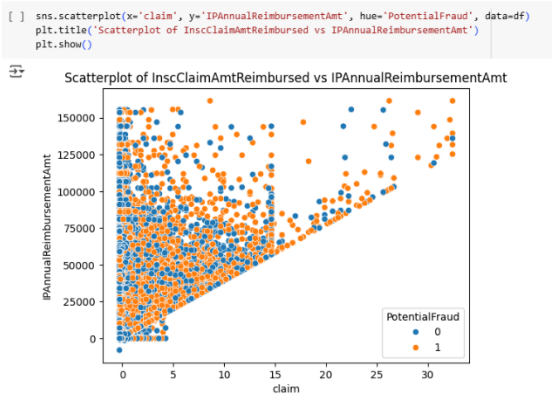

Fig. 2. Scatter Plot of Claims vs. Reimbursement Amount

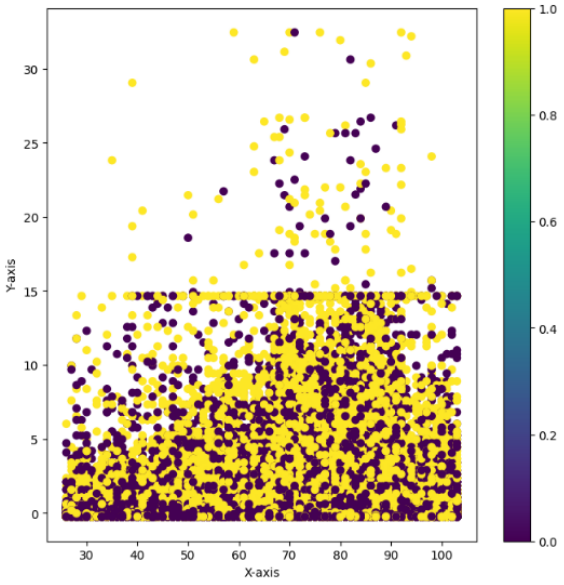

Fig. 3. Scatter Plot: Age vs. Claim Amount.

The scatter plot helps in identifying patterns of fraud in claim amounts by plotting the correlation of insurance claim amounts and fraudulent activities. The horizontal axis in the graph represents claim amounts and the vertical axis represents the fraud classification where 1 shows the fraudulent claims while 0 is for non-fraudulent claims. The line at y=1 represents the claims marked as fraudulent and the line at y=0 is for non-fraudulent claims. The distribution shows that it is not limited to specific claim amount and higher claim amount is more likely to be a fraudulent claim. This means that any claim that is large should attract more attention as it will always display certain features that are likely to make one think that it is fraudulent.

For this code we use MinMax Scaling in order to scale all numerical features to the range of between 0 to 1 using MinMaxScaler(). This makes it a prerequisite that all the features have an equal range so as to minimize the impact of large values on the training of the model. SelectKBest(chi2, k=4) is used to select the best features and the Chi-Square test is used to select the top four most relevant features that can be used in the prediction of fraudulent providers. This way, the model's accuracy increases, but computational resources needed for its training are decreased since only the most informative features are selected. The last dataset has 542,151 rows and 4 chosen columns.

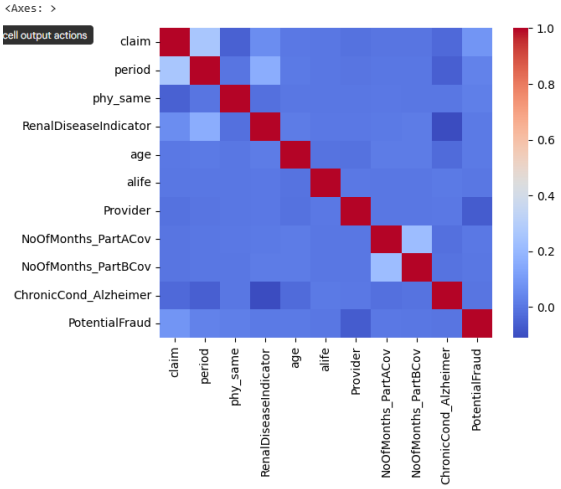

Fig. 4. Correlation Analysis for Fraud Detection.

The heatmap provides an insight of how the Potential Fraud column is related to the other features of the dataset. The values of correlation coefficients are 0.178 for fraud occurrences and the selected features, which means that there is a very low level of association. This implies that it may be unadvisable to use regression-based models for fraud detection since they are based on high degrees of relationships between the variables. This is why, there is no linear relationship and, therefore, fraud detection needs to use decision trees, gradient boosting, or neural networks. The future work can also be done in the direction of feature engineering to find more relevant features that can be used to predict the fraud cases definitely.

In this code the data set is divided into 70% training data set and 30% is for testing data set using train_test_split function from sklearn.model_selection. The set (X_train, y_train) is used for training of the model and the set (X_test, y_test) is used to assess the performance of the proposed model. The parameter random_state=42 makes the model consistent by generating the same results every time the program is run.

This code fits a Decision Tree Classifier by using DecisionTreeClassifier() from the sklearn.tree module. The parameter max_depth=6 restricts the tree to 6 levels in order to avoid extensive predictions while ensuring interpretability. The model fit is applied using Xtrain and Ytrain as the fitting parameters. As a result, the classifier predicts the outcomes for Xtest_selected, which is a subset of Xtest.

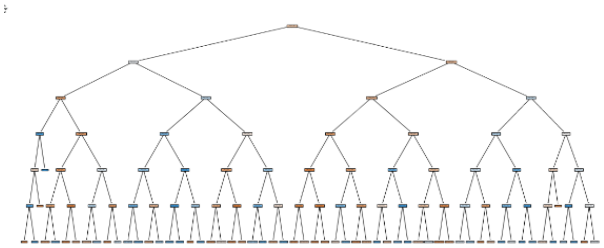

Fig. 5. Decision Tree.

Fig. 6. Model Performance by Score.

The graph shows the performance of various machine learning algorithms that can be used for fraud detection in terms of accuracy, precision, recall, and F1-score. Neural Networks give the highest value of recall – this means that the Neural Networks are likely to detect most of the fraud cases but this is at the expense of the precision. GBM with hyperparameter tuning is the most effective, as it is consistently good for all the evaluation criteria.

4 Discussion

The evidence presented in this study further supports using Retrieval-Augmented Generation (RAG) system in the contextualization and reuse of composed structured scientific datasets, especially in critical domains such as health. Real-time retrieval with generative functions improves the contextual use and the resulting output. Nevertheless, these benefits come with ethical and technical issues that has to be solved through proper design. Another one of the findings got from this study is on the suitability of privacy-preserving techniques including federated learning and homomorphic encryption. These methods enable the distributed data processing and minimize the threat of the centralized attacks and meet todays' data protection standards (Koga et al., 2024; Geren et al., 2024). Further, this study calls for the incorporation of interpretability and traceability features when used in the RAG system, especially if it discerns the end-user in sensitive areas. Explanations such as decision trees or SHAP values heatmaps

are useful for increasing trust and making users explain their concerns regarding credibility and accountability (Labajová, 2023).

The result of analyzing the results obtained from the experiment tends to support the hypothesis that non-linear model is more effective than the linear models in the detection of fraud. Thus, the models like XGBoost and Neural Networks proved to be more effective in the key metrics like recall and accuracy and are ideal for cases when less frequently occurring but highly expensive frauds are present. However, the interpretability of results must not be ignored at the cost of model performance which tree-based models easily meet in the regulatory framework. The nature of the dataset is revealed by the nearly normal distribution of fraud counts, and this calls for the use of resampling methods and also evaluation metrics that do not overemphasize the majority class.

In addition, this research pays attention to strengthen RAG systems against adversarial threats where the adversarial perturbation or poisoning attacks can affect the retrieval or generation process (Nazary et al., 2025; Chen et al., 2025). The security of these models can be bolstered by adversarial training and also the ways the data is ingested securely. From a DHC perspective, this study is related to the principles of sustainable and smart curation by providing approaches to the secure and ethical reusing of the structure data. When using healthcare data for model training and fraud detection, it is evident that AI systems can try to find ways to utilize secondary uses of data while adhering to legal and ethical standards. These ideas are relevant to other industries such as cultural heritage centers, schools and libraries where reliable artificial intelligence is necessary in today's digital world.

Application of RAG in Cultural and Scientific Heritage

While this study primarily explores the application of Retrieval-Augmented Generation (RAG) systems in healthcare, these models also hold significant promise for the curation and dissemination of cultural and scientific heritage. Institutions such as museums, archives, and libraries house extensive collections of textual and multimedia content that are often underutilized due to their complexity and language diversity. RAG can facilitate semantic enrichment and multilingual generation of archival descriptions, automate metadata extraction, and support cross-lingual search capabilities across cultural databases (Gupta et al., 2024). For instance, museum curators could use RAG to generate inclusive exhibit summaries based on diverse historical sources. In the realm of scientific heritage, RAG can synthesize scholarly outputs for research continuity and knowledge transfer in disciplines like archaeology and philology (Lavingia & Mehta, 2022). Importantly, when integrated with privacy-preserving techniques such as differential privacy and federated learning (Koga et al., 2024), these systems can ethically process sensitive cultural data, including indigenous or restricted-access archives. As concerns around trust and bias in generative AI persist (Labajová, 2023), embedding interpretability tools into RAG pipelines becomes vital for accountability. Thus, beyond healthcare, RAG systems offer transformative potential in ensuring accessible, secure, and context-aware engagement with cultural and scientific knowledge.

5 Conclusions

Retrieval augmentation (RAG) is a significant concept for security, ethical, and intelligent reprocessing of sensitive scientific data. The authors in this work demonstrated how big-scale health care data can be semantically cleaned and analyzed using machine learning algorithms with data privacy constraints imposed. In this way, it is possible to include federated learning and interpretability tools in RAG as such: In the future researches, the authors plan to conduct more advanced experiments with RAG, for example, using it for fraud detection in the stream, incorporating more various languages into both resourceful and scarce corpora etc. The following are measures that are aimed at furthering RAG's involvement in the stewardship of data in the areas of health, education, and other digital cultural domains (Jiang et al., 2024).

Acknowledgements.

My appreciation goes to all those who have supported the accomplishment of the research work. First of all, it is necessary to thank the aca-demic supervisor at SRH Hochschule Berlin for the feedback and assistance provided while completing the project. They contributed a lot in guiding the contextual direction of this work. I also extend my appreciation to all the faculty and MSc Computer Science students for their cooperation and contribution in improving my research experience. We all thank the developers and contributors of the used Python packages such as scikit-learn, pan-das, seaborn for enabling us process the data and also for enabling us present the data in this form.

I would like to highly appreciate the authors of the publicly available healthcare fraud detection dataset from which the data used in this study has been extracted. They are essential in building the enhancement of the academic research in the case of the open data projects. Finally, I would like to thank my family and friends because their encouragement compensated for the tangible and constant support during the process of research.

References

Aly, A., & Smart, N. P. (2019, May). Benchmarking privacy-preserving scientific operations. In *International Conference on Applied Cryptography and Network Security* (pp. 509–529). Springer International Publishing. https://eprint.iacr.org/2019/354.pdf

Asai, A., Min, S., Zhong, Z., & Chen, D. (2023, July). Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)* (pp. 41–46). https://aclanthology.org/2023.acl-tutorials.6.pdf

Chen, Z., Gong, Y., Chen, M., Liu, H., Cheng, Q., Zhang, F., Lu, W., Liu, X., & Liu, J. (2025). FlipedRAG: Black-box opinion manipulation attacks to retrieval-augmented generation of large language models. *arXiv preprint* arXiv:2501.02968. https://arxiv.org/pdf/2501.02968

Geren, C., Board, A., Dagher, G. G., Andersen, T., & Zhuang, J. (2024). Blockchain for large language model security and safety: A holistic survey. *arXiv preprint* arXiv:2407.20181. https://arxiv.org/pdf/2407.20181

Gupta, S., Ranjan, R., & Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions. *arXiv preprint* arXiv:2410.12837. https://arxiv.org/pdf/2410.12837

Jiang, C., Pan, X., Hong, G., Bao, C., & Yang, M. (2024). Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint* arXiv:2411.14110. https://arxiv.org/pdf/2411.14110

Koga, T., Wu, R., & Chaudhuri, K. (2024). Privacy-preserving retrieval augmented generation with differential privacy. *arXiv preprint* arXiv:2412.04697. https://arxiv.org/pdf/2412.04697

Labajová, L. (2023). The state of AI: Exploring the perceptions, credibility, and trustworthiness of users towards AI-generated content. https://www.diva-portal.org/smash/get/diva2:1772553/FULLTEXT02

Lavingia, K. R., & Mehta, R. (2022). Information retrieval and data analytics in Internet of Things: Current perspective, applications and challenges. *Scalable Computing: Practice and Experience, 23*(1), 23–34. https://scpe.org/index.php/scpe/article/download/1969/716

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, 1–28. https://link.springer.com/content/pdf/10.1007/s43681-024-00427-4.pdf

Nazary, F., Deldjoo, Y., & di Noia, T. (2025). Poison-RAG: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. *arXiv preprint* arXiv:2501.11759. https://arxiv.org/pdf/2501.11759

Received: April 15, 2025
Reviewed: May 05, 2025
Finally Accepted: May 15, 2025