AI-Ready Multimodal Data Pipeline to Enrich Cancer Care

Rezaur Rashid[1*], Soheil Hashtarkhani[1], Fekede Asefa Kumsa[1], Lokesh Chinthala[1], Brianna M White[1], Janet A ZinK[1], Christopher L Brett[2], Robert L Davis[1], David L Schwartz[1*], Arash Shaban-Nejad[1*]

[1] *College of Medicine, The University of Tennessee Health Science Center, Memphis, TN, USA*
[2] *The University of Tennessee Graduate School of Medicine, Knoxville, TN, USA*

Corresponding authors:*
Rezaur Rashid (*mrashid7@uthsc.edu),* David L Schwartz (*dschwar4@uthsc.edu*), Arash Shaban-Nejad (*ashabann@uthsc.edu*)
College of Medicine, The University of Tennessee Health Science Center, Memphis, TN, USA

2024 IEEE International Conference on Big Data (BigData) | USGov | DOI: 10.1109/BigData62323.2024.10825353

***Abstract*— This study reports on the progress in designing and developing a framework for integrating heterogeneous datasets—structured, semi-structured, and unstructured—into an AI-ready multimodal data pipeline aimed at predicting radiation therapy interruptions (RTI) and enhancing patient care navigation. The AI-Ready dataset incorporates a broad set of information, including patient demographics, health data, clinical notes, medical imaging, and data on social determinants of health. Preliminary results indicate that this pipeline effectively integrates diverse distributed data sources, providing a foundation for training AI models capable of generating reliable and actionable predictions.**

I. Introduction

Advances in artificial intelligence (AI) are transforming the healthcare paradigm, introducing predictive models that enhance clinical decision-making and improve patient outcomes. Notably, in the domain of cancer care, radiation oncology faces significant challenges posed by radiation treatment interruptions (RTIs), which can compromise treatment efficacy and adversely affect patient survival rates [1, 2]. RTIs increased during the COVID-19 pandemic, especially in underserved areas where limited healthcare access and resources made it harder to maintain consistent cancer treatments [3]. To mitigate this issue, predictive modeling of RTIs becomes essential, requiring the integration of diverse healthcare data sources to comprehensively capture the complexities of patient-specific circumstances [4, 5].

Healthcare data's multimodal nature encompasses structured Electronic Health Records (EHR), semi-structured clinical notes, and unstructured imaging data. Integrating these diverse data types into a unified AI-ready format poses significant challenges driven by heterogeneity in formats, inconsistent data quality, and the need for scalable reproducible solutions [6]. Existing research often focuses on single-modality approaches, which limits predictions by lowering accuracy and missing important insights that can be gained from integrated data sources [7].

Recent studies demonstrate incorporating multiple data types enhances clinical predictions [8, 9]. However, challenges exist in harmonizing structured data, text, and medical images into a consistent dataset. In addition, in some cases, privacy concerns may require synthetic data generation to ensure patient confidentiality [10, 11].

This research was supported by a grant from the Tennessee Department of Health.

In a clinical and public health context, AI-readiness refers to the quality and preparation of diverse (health and non-health) datasets, ensuring they are suitable for artificial intelligence (AI) applications to generate targeted, actionable insights that support patient care, improve health outcomes, and inform population-level health interventions. This study develops an AI-ready multimodal data pipeline for RTI prediction, addressing key challenges such as data fragmentation and variable data quality to enable more accurate and scalable predictive models while integrating a diverse data type into a single, cohesive dataset. Our key contributions include:

- Design and development of an AI-ready multimodal data integration pipeline unifying structured, EHRs, clinical text, and imaging data for predictive modeling in radiation oncology.
- Implementation of a unified data representation enhancing model efficiency while maintaining clinical interpretability.
- Case study application showcasing the pipeline's effectiveness in improving RTI prediction accuracy through multimodal data integration.

This paper aims to tackle the challenges of creating AI-ready datasets by introducing a comprehensive end-to-end pipeline that integrates diverse multimodal healthcare data. The following sections detail the methodology, experiments, and results, demonstrating the pipeline's ability to advance AI model performance while maintaining clinical relevance and applicability.

II. Related Work

The quest for AI-ready healthcare data has intensified, driven by the potential of AI to enhance clinical decision-making and patient outcomes. Most current approaches focus on preparing individual data modalities, such as structured EHR, clinical notes, or medical imaging, but often fail to fully integrate these disparate data types into a unified dataset that is suitable for machine learning. Willemink et al. [12] addressed critical challenges in data preprocessing, particularly in the context of medical imaging, establishing methods for creating high-quality, AI-ready imaging datasets. Diaz et al. [13] expanded on this, providing a comprehensive guide for preparing imaging data using open-access platforms and harmonization techniques to enhance data quality and interoperability. However, these studies generally focus on a

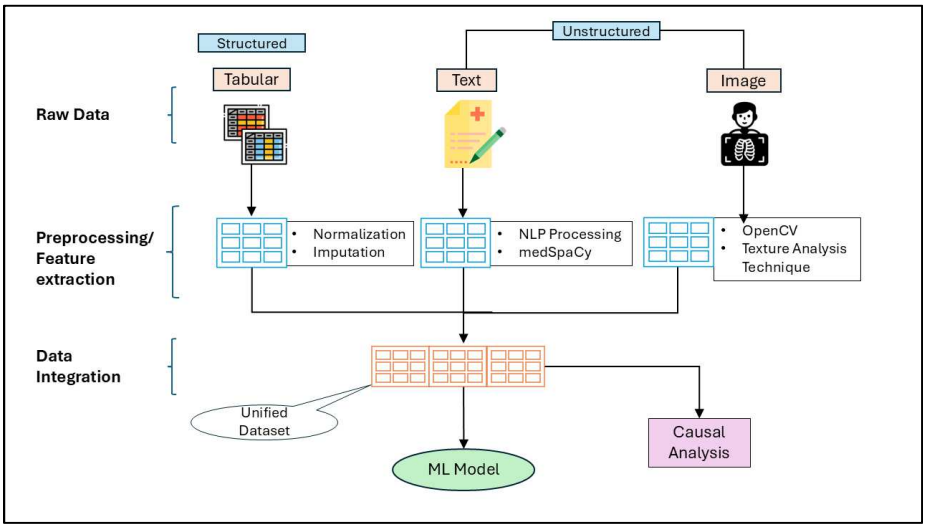

Fig. 1. AI-Ready Multimodal Data Pipeline

single modality data and lack an integrated view that combines structured EHR data with unstructured clinical notes and imaging data.

Recent studies have demonstrated the value of integrating multiple data types to enhance predictive performance [14, 15]. For example, Miotto et al. [8] employed unsupervised deep learning to create patient representations from structured EHR data, improving patient outcome predictions. Similarly, Rajkomar et al. [9] used deep learning to process structured healthcare data, highlighting the potential for EHR-based predictions. Despite these advances, both studies did not integrate unstructured data, limiting their ability to capture the full context necessary for comprehensive patient modeling.

Natural Language Processing (NLP) techniques, such as BERT-based and SpaCy-based models, have improved the extraction of clinically relevant entities from textual notes [16, 17]. For example, Alsentzer et al. [16] leveraged BERT-based language models to extract patient details from clinical texts, offering improved representational capabilities compared to traditional NLP methods. However, while NLP advancements have improved the extraction of meaningful information from clinical notes, these approaches often fail to fully integrate textual features with structured EHR or unstructured imaging data, which is critical for a holistic understanding of patient conditions. Similarly, imaging data integration has advanced significantly, primarily with convolutional neural networks (CNNs) such as ResNet for feature extraction. Xu et al. [18] demonstrated the effectiveness of combining imaging features with clinical data, showing improved performance in diagnostic tasks. Yet, these models often neglect unstructured clinical notes, missing important contextual information about patient history and ongoing treatment that could improve prediction outcomes.

The challenge of creating an integrated, multimodal dataset is further complicated by privacy concerns and data silos, which restrict data availability. Thomas et al. [19] highlighted the need for transparent, fair, and integrated health data for AI applications, , while addressing the complexities of fragmented healthcare data.

Our work addresses these limitations by proposing an AI-ready data pipeline that integrates structured data (such as patient demographics, treatment plans, and social determinants of health (SDoH) metrics), clinical notes, and imaging features into a unified dataset. Unlike prior studies, we prioritize interpretability by focusing on categorical feature extraction instead of embeddings, ensuring that the model outputs are not only accurate but also explainable for clinicians. Our methodology bridges the gaps between different data modalities offering a comprehensive approach to creating AI-ready datasets for cancer care, with a specific focus on Radiation Treatment Interruption (RTI) prediction in radiation oncology.

III. Methodology

The goal of this study is to design and develop an AI-ready data pipeline for addressing the challenges of data fragmentation and multimodality to accurately predict Radiation Treatment Interruptions (RTI).

A. AI-Ready Data Pipeline: Overview

The AI-ready data pipeline is designed to systematically integrate multimodal healthcare data, ensuring that the final dataset is not only cohesive, and interpretable but also suitable for machine learning applications. The foucus was on extracting clinically relevant, interpretable categorical features rather than dense embeddings, to enhance explainability and support causal analysis. This approach aims to provide clinicians with clear, actionable insights that directly inform patient care.

The pipeline is composed of a series of interconnected steps, including data ingestion, processing, integration, and preparation for predictive modeling. Fig. 1 demonstrates a flowchart of the complete pipeline, detailing the transformation of raw healthcare data into a unified format optimized for AI modeling.

B. Data Acquisition and Sources

Structured Data. Demographics, clinical metrics, and SDoH metrics were collected from 1,840 patients. Structured data included features such as age, gender, marital status, treatment intent, and distance to treatment facilities.

Text-Based Data. A total of 5,925 clinical notes and treatment plans in PDF format were collected, containing valuable contextual information about patient conditions, treatment strategies, and reported side effects.

Imaging Data. To evaluate the pipeline's multimodal integration capabilities, chest X-ray (CXR) data was obtained from MIMIC-CXR [20]. Although this dataset was not originally part of the primary dataset, they were included to illustrate the feasibility of integrating imaging data with other data modalities for downstream predictive analysis.

C. Data Preprocessing and Feature Extraction

a) Structured Data Extraction: Demographics and clinical metrics were extracted, standardized, and preprocessed using pandas and Scikit-learn. To ensure data consitency, missing values were imputed using mean imputation for numerical features and mode imputation for categorical features.

b) Text Data Extraction: Clinical notes and treatment plans were processed using medSpaCy [21], a alspecialized NLP toolkit for medical data. The tool extracted entities that were relevant to RTI predictions, including patient conditions (e.g., lung cancer), treatment types (e.g., chemotherapy), and side effects (e.g., nausea, fatigue). Instead of creating text embeddings, the pipeline focused on extracting interpretable categorical features to maintain explainability and clinical relevance.

c) Image Data Extraction: Chest X-ray images were preprocessed using OpenCV for noise reduction, followed by feature extraction using texture analysis and shape-based techniques (e.g. Gray Level Co-occurrence Matrix (GLCM)). Unlike deep learning approaches that generate dense embeddings, we extracted higher-level features that could be easily categorized and interpreted by clinicians. Features such as the presence/absence of abnormalities, lesion size, and anatomical location were captured, allowing us to create meaningful and interpretable image-derived features.

d) Motivation for Categorical Feature Extraction: Given the emphasis on explainable AI (XAI), categorical features were prioritized over embeddings for both text and image data. While embeddings are effective at capturing information, often lack interpretability. Instead, features like tumor presence, disease categories, and treatment types were selected to ensure alignment with clinical interpretability requirements. These features make the AI outputs actionable for clinicians by providing a basis of interpretability and relevance in potential causal pathways., unlike embeddings, which are more abstract.

D. Data Integration

Once features were extracted from each modality, they were integrated into a unified dataset. Key steps in the integration process included:

- **Common Identifier**: Data from different sources (structured, text, imaging) were linked using a common patient identifier to ensure data consistency across modalities.

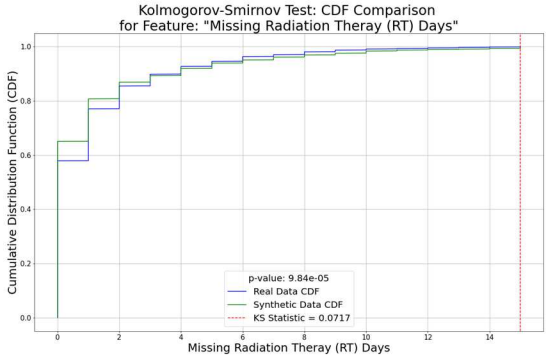

Fig. 2. KS-statistic Comparison between Real Data and Synthetic Data

- **Feature Alignment**: Features from each modality were aligned, and categorical attributes were encoded as needed. Label encoding for categorical variables was used instead of one-hot encoding to preserve interpretability, minimizing feature space expansion.
- **Validation**: The pipeline was validated to ensure all features were interpretable and suitable for predictive modeling. Consistency checks were conducted to verify alignment across patient records, ensuring completeness and feasibility of downstream analysis.

IV. Experiments

The experimental phase aimed to validate the effectiveness of the AI-ready multimodal data pipeline through two case studies each focusing on different modalities and integration challenges. **Case Study 1** involved synthetic data derived from real patient records, focusing on integrating structured data with textual features. In contrast, **Case Study 2** focused on the MIMIC-CXR dataset to demonstrate how imaging data can be integrated with textual data. The purpose of these experiments was to evaluate the pipeline's capability to process and integrate multiple data modalities for machine learning tasks.

A. Case Study 1: Synthetic Data from Real Patient Records

The original dataset contained information from 1,840 patients and 5,925 clinical treatment plans in PDF format. This data included structured elements such as patient demographics, treatment plans, and SDoH, alongside unstructured features extracted from clinical notes. Given the limited sample size and privacy concerns, synthetic data generation was employed to create a dataset that preserves patient confidentiality while enabling model training.

We used CTGAN (Conditional Tabular Generative Adversarial Network) [11], an extension of GANs optimized for tabular data, to generate synthetic records. CTGAN was chosen for its ability to accurately model the complex relationships and mixed data types commonly found in healthcare datasets. The generator and discriminator network within CTGAN learned the joint distribution of the structured dataset to produce realistic synthetic samples. The quality of the synthetic data was evaluated using metrics such as the Kolmogorov-Smirnov (KS) statistic, ensuring the generated feature distributions closely

resembled the original data (Fig. 2). The final synthetic dataset consisted of 5,000 patient records, which served as the training and validation data for the RTI prediction model.

Although the synthetic data generated by CTGAN effectively mirrors the statistical properties and relationships found in real-world data, it may still lack the complete variability inherent in actual clinical environments. Therefore, these synthetic datasets serve as a strong basis for model development, but validation on larger real-world datasets remains a necessary future step to confirm their robustness and applicability.

B. Case Study 2: MIMIC-CXR Dataset

To explore the integration of imaging data within the pipeline, we used MIMIC-CXR, containing chest X-ray images with corresponding radiology reports. We selected and processed 5,000 chest X-ray images. The textual notes were extracted similarly to the clinical notes in the synthetic dataset from Case Study 1, ensuring consistency in feature extraction. This case study illustrated the pipeline's potential to create a more comprehensive patient assessment by integrating imaging features with textual information.

It is important to clarify that the imaging data in Case Study 2 served as a proof of concept to demonstrate the versatility of our AI-ready data pipeline for multimodal integration. It was independent of the synthetic data from Case Study 1, focusing not on improving prediction accuracy but on showcasing the pipeline's capability to handle diverse data types.

C. Prediction Task and Model Training

The goal of both case studies was formulated as a classification problem with specific prediction tasks:

- In **Case Study 1**, structured and text-based features were used for a binary classification. The target variable was defined as 'Yes' if a patient had two or more missed treatment days (classified as "RTI") and 'No' otherwise.

- In **Case Study 2**, combined features from text and imaging data were used for a multi-level classification task, aiming to classify patients into different disease severity levels based on the findings in the X-ray images.

In both case studies, XGBoost, a Gradient Boosting classifier, was selected for predictive model training due to its robustness and durability for tabular health data. XGBoost was configured with hyperparameter optimization through a grid search, tuning the learning rate, maximum depth, and number of estimators to determine the optimal parameters for model performance.

D. Model Evaluation

The models were evaluated using metrics such as Accuracy, Area Under the Curve (AUC), and F1-score. For Case Study 2, we calculated both one-vs-rest (OVR) and one-vs-one (OVO) AUC scores to better capture the performance in the multi-class scenario. These metrics provided a comprehensive assessment of model performance, emphasizing both classification correctness and the balance between sensitivity and specificity, critical factors for clinical decision-making.

V. Results

The section presents a comparative evaluation across both case studies, highlighting the impact of different combinations of data modalities on model performance.

A. Performance Metrics for Case Study 1 (Synthetic Data)

Table I presents the results for **Case Study 1**, where models were trained using structured data only, text data only, and a combination of both structured and text-derived features. The combination of structured and text-derived features consistently outperformed models trained on individual modalities, achieving higher AUC, Accuracy, and F1-score. This improvement underscores the critical role of contextual information embedded within clinical notes, such as treatment types and potential side effects in enhancing the model's ability to predict RTI.

TABLE I. Model Performance for Different Data Modalities in Case Study 1 (Synthetic Data)

Data Type	AUC	Accuracy	F1-score
Structured only	0.819	76.00%	0.707
Text only	0.593	63.50%	0.330
Structured + Text	**0.821**	**78.25%**	**0.727**

B. Performance Metrics for Case Study 2 (MIMIC-CXR)

Table II presents the performance metrics for Case Study 2, where models were trained using text-only, image-only, and combined text and image-derived features. The findings suggest that: Text-Only Features outperformed image-only features, underscoring the significant predictive value embedded in radiology reports. Image-only features underperformed, indicating that, without the contextual understanding provided by textual data, imaging data alone was insufficient in this scenario.

Interestingly, combining image features with text features led to a decrease in AUC compared to using text alone. This may suggest that the image data did not add complementary value to this prediction task but rather introduced noise, possibly because of differences in the timing and context of image acquisition relative to the radiology reports.

TABLE II. MODEL PERFORMANCE FOR DIFFERENT DATA MODALITIES (CASE STUDY 2).

Data Type	AUC (OVR)	AUC (OVO)	Accuracy	F1-score
Text only	**0.861**	**0.823**	**66.98%**	**0.662**
Image only	0.539	0.541	37.00%	0.205
Text + Image	0.845	0.808	65.57%	0.646

VI. DISCUSSION

A. Findings

Integration text-derived features consistently improved model performance, underscoring the value of unstructured clinical notes. In Case Study 1, text-derived features enhanced

performance compared to structured data alone, highlighting the role of contextual information from clinical notes.

In Case Study 2, text data provided strong predictive power, while imaging-only features performed poorly, suggesting that imaging data without context has limited utility. These findings emphasize that the relevance and effectiveness of imaging data are highly dependent on its clinical context and the quality of integration with other data types.

B. Challenges in Synthetic Data Generation

Ensuring the statistical integrity of synthetic data posed significant challenges, particularly in maintaining the fidelity of complex feature distributions. To address this, rigorous monitoring of feature distributions and bias mitigation were conducted using the Kolmogorov-Smirnov (KS) statistic and correlation analysis. These measures ensured that the synthetic dataset closely resembled the real-world data, supporting reliable model training while preserving privacy.

C. Limitations in Feature Extraction and Integration

The extraction of categorical features from text and images was prioritized for interpretability, a key requirement for clinical applications. However, this process required significant effort to align features across modalities. Focusing on categorical features instead of embeddings supported clinical decision-making by enhancing explainability but introduced challenges in maintaining consistency across data types. This limitation highlights the trade-off between interpretability and the uniformity of feature representation.

D. Data Availability

The syntactic data used in this study are available upon reasonable request.

VII. Conclusion

This work presents early work on the design and development of a pipeline for integrating heterogeneous healthcare data into an AI-ready dataset to study radiation oncology treatment interruptions. The pipeline's utility and effectiveness were validated through two case studies, demonstrating that integrating multiple data modalities improves model performance. This finding emphasizes the importance of comprehensive data pipelines in healthcare for supporting advanced predictive models.

By focusing on the extraction of clinically interpretable categorical features and ensuring the dataset's readiness for AI applications, this study lays the groundwork for future machine learning models that are not only accurate but also actionable in clinical settings. Future Work will expand the dataset to include more diverse patient populations, incorporate additional methods for causal analysis, and enhance model interpretability. These enhancements will ensure that the resulting AI model is both robust and readily adoptable by clinicians.

References

[1] T. Shaikh, et al., "The impact of radiation treatment time on survival in patients with head and neck cancer," International Journal of Radiation Oncology* Biology* Physics, vol. 96, no. 5, pp. 967–975, 2016.

[2] A. Shaban-Nejad, et al., "Towards an explainable ai platform to study interruptions in cancer radiation therapy," in MEDINFO 2023—The Future Is Accessible, pp. 1501–1502, IOS Press, 2024.

[3] E. Gaudio, et al., "Defining radiation treatment interruption rates during the covid-19 pandemic: findings from an academic center in an underserved urban setting," International Journal of Radiation Oncology* Biology* Physics, vol. 116, no. 2, pp. 379–393, 2023.

[4] A. Joseph, T. Hijal, J. Kildea, L. Hendren, and D. Herrera, "Predicting waiting times in radiation oncology using machine learning," in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1024–1029, IEEE, 2017.

[5] M. Martínez-García and E. Hernández-Lemus, "Data integration challenges for machine learning in precision medicine," Frontiers in medicine, vol. 8, p. 784455, 2022.

[6] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," Proceedings of the IEEE, vol. 103, no. 9, pp. 1449–1477, 2015.

[7] H. Uno, D. P. Ritzwoller, A. M. Cronin, N. M. Carroll, M. C. Hornbrook, and M. J. Hassett, "Determining the time of cancer recurrence using claims or electronic medical record data," JCO clinical cancer informatics, vol. 2, pp. 1–10, 2018.

[8] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," Scientific reports, vol. 6, no. 1, pp. 1–10, 2016.

[9] A. Rajkomar, et al., "Scalable and accurate deep learning with electronic health records," NPJ digital medicine, vol. 1, no. 1, pp. 1–10, 2018.

[10] K. Armanious, C. Jiang, M. Fischer, T. K¨ustner, et al.,, "Medgan: Medical image translation using gans," Computerized medical imaging and graphics, vol. 79, p. 101684, 2020.

[11] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," Advances in neural information processing systems, vol. 32, 2019.

[12] M. J. Willemink, et al., "Preparing medical imaging data for machine learning," Radiology, vol. 295, no. 1, pp. 4–15, 2020.

[13] O. Diaz, et al., "Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools," Physica medica, vol. 83, pp. 25–37, 2021.

[14] Y. Yang, K. Sun, Y. Gao, K. Wang, and G. Yu, "Preparing data for artificial intelligence in pathology with clinical-grade performance," Diagnostics, vol. 13, no. 19, p. 3115, 2023.

[15] F. Kidwai-Khan, R. Wang, M. Skanderson, C. A. Brandt, S. Fodeh, and J. A. Womack, "A roadmap to artificial intelligence (ai): Methods for designing and building ai ready data to promote fairness," Journal of Biomedical Informatics, vol. 154, p. 104654, 2024.

[16] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," arXiv preprint arXiv:1904.03323, 2019.

[17] J. Lee, et al., "Biobert: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

[18] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," Journal of healthcare informatics research, vol. 5, pp. 1–19, 2021.

[19] D. M. Thomas, et al., "Transforming big data into ai-ready data for nutrition and obesity research," Obesity, vol. 32, no. 5, pp. 857–870, 2024.

[20] A. E. Johnson, et al., "Mimic-cxr, a deidentified publicly available database of chest radiographs with free-text reports," Scientific data, vol. 6, no. 1, p. 317, 2019.

[21] H. Eyre, et al., "Launching into clinical space with medspacy: a new clinical text processing toolkit in python," in AMIA Annual Symposium Proceedings, vol. 2021, p. 438, 2022.