Annual Report (2024–2025)

Radiomics-Based Tumor Phenotypes to Predict Individual Risk and Treatment Response in Head and Neck Cancer

- **Funding Agency:** DBT Wellcome Trust India-Alliance
- **Project Duration:** 1 January 2020 – 31 December 2025
- **Funding Amount:** ₹1,35,13,928
- **Principal Investigator (PI):** Dr. Hannah Mary Thomas T
- **Project Assistant (PA):** Mr. Hasan Shaikh
- **Department:** Radiation Oncology, Unit II, Christian Medical College, Vellore

Background and Significance

Cancer radiomics is an emerging field in medical imaging that focuses on converting routine radiological images—traditionally interpreted qualitatively—into quantifiable data describing tumor phenotypes. When combined with advanced statistical analytics, radiomics can significantly improve the accuracy of clinical outcome prediction models.

Radiomics operates on the principle that imaging characteristics such as texture and intensity reflect underlying gene-protein expressions or tumor phenotypes. Unlike tissue biopsies, radiomics captures information from the entire tumor, reducing sampling errors. This imaging data is obtained non-invasively from routine clinical scans, requiring no additional procedures or radiation exposure for patients.

Importance of the Proposed Research:

Accurately estimating an individual cancer patient's risk of early disease failure is crucial for:

1. Understanding tumor biology
2. Stratifying patients based on risk
3. Tailoring personalized treatment strategies
4. Optimizing the use of limited radiotherapy resources

This project aims to study quantitative imaging parameters (radiomic features) from pretreatment CT and PET scans in head and neck cancer (HNC) patients. Developing robust image analysis pipelines, feature extraction tools, and predictive models will enable us to answer a key research question: **Can pre-treatment radiomics signatures accurately identify advanced HNC patients at higher risk of recurrence and poor survival outcomes?**

Specific Aims of the Project:

1. **Develop a robust tumor segmentation and radiomic feature extraction pipeline for PET and CT imaging in Head and Neck Cancer patients.**

2. **Build predictive models for head and neck cancer patients using pre-treatment radiomic features to estimate patient outcomes.**

3. **Design a comprehensive Head and Neck Cancer imaging archive for validation, clinical translation, and training on the radiomics-based risk stratification platform.**

Progress During Reporting Period (September 2024 – March 2025)

Aim 1: Develop a robust tumor segmentation and radiomic feature extraction pipeline for PET and CT imaging in Head and Neck Cancer patients.

During this period, significant progress was made in developing an automated tumor segmentation workflow using anonymized CT imaging data of Head and Neck Cancer (HNC) patients from our institutional dataset. As part of a collaboration with the Department of Computer Science and Engineering at the National Institute of Technology Karnataka (NITK), Surathkal, anonymized patient datasets were shared to support the development and evaluation of a deep learning-based segmentation model based on the 3D nnU-Net framework. The model focused on the fully automated delineation of the primary gross tumor volume (GTV) from CT scans, aiming to enhance consistency, accuracy, and efficiency in radiotherapy planning. The collaborative work on automated segmentation led to the submission of the manuscript titled **"Automated Segmentation of Head and Neck Cancer from CT Images Using 3D Convolutional Neural Networks"** to the *Journal of Imaging Informatics in Medicine (JIIM)* for peer review.

Aim 2: Build predictive models for head and neck cancer patients using pre-treatment radiomic features to estimate patient outcomes.

Radiomics-based predictive models were developed by applying various feature selection and optimization strategies followed by machine learning classifiers.

Feature Selection and Optimization Methods Implemented:

- Least Absolute Shrinkage and Selection Operator (LASSO): Regularization-based feature selection.
- SelectKBest: Univariate statistical feature selection.
- Particle Swarm Optimization (PSO) & Whale Optimization Algorithm (WOA): Swarm intelligence-based feature optimization.

Classifiers Used: These feature-selected datasets were utilized to train and evaluate the following classifiers:

- Logistic Regression
- Naive Bayes Classifier
- Linear Support Vector Machine (Linear SVM)
- Radial Basis Function Kernel Support Vector Machine (RBF SVM)
- Decision Tree Classifier
- Random Forest Classifier
- Voting Classifier (ensemble approach combining multiple models)

Each model's performance was evaluated primarily using ROC AUC and accuracy metrics to determine their predictive capability for locoregional recurrence in head and neck cancer patients.

Aim 3: Design a comprehensive Head and Neck Cancer imaging archive for validation, clinical translation, and training on the radiomics-based risk stratification platform.

- **Data Collection Status:** A total of **1550 HNC patients' data** have been collected prospectively, forming a valuable imaging archive for validation and future studies.

Project Staff Contributions (September 2024 – March 2025)

Mr. Hasan Shaikh, Project Assistant, Joined the project in September 2024. His achievements during this reporting period include:

Workshops, Seminars, and Symposiums Attended:

- Participated in the **14th Annual Research Day**, Office of Research Christian Medical College, Vellore (October 24–25, 2024).
- Attended the **Continuing Medical Education (CME)** program on *"Revolution and Precision in Radiation Oncology,"* Ida B. Scudder Cancer Center, CMC (March 1, 2025).
- Participated and served as part of the Organizing Team for the **2nd Annual Winter Symposium on Health Data and AI**, Biomedical Informatics Unit, CMC Vellore (March 13–15, 2025).

Oral Presentations:

- Hasan Shaikh, Amal Joseph Varghese, Hannah Mary Thomas T, et al., *"Can CT Radiomics Predict Recurrence in Head and Neck Cancer? Early Results from a Prospective Imaging Trial,"* presented at the 14th Annual Research Day, CMC Vellore (October 24–25, 2024).
- Piyus Prabhanjans, Aparna V. K., Rajendra Benny Kuchipudi, Hannah Mary Thomas T, Balu Krishna S, Hasan Shaikh, Amal Joseph Varghese, Simon Pavamani, and Jeny Rajan, *"Automated Segmentation of Head and Neck Cancer from CT Images Using 3D Convolutional Neural Networks,"* presented at the 2nd Annual Winter Symposium on Health Data and AI (March 13–15, 2025).

Publications:

- Manuscript submitted: *Piyus Prabhanjans, Aparna V. K., Rajendra Benny Kuchipudi, Hannah Mary Thomas T, Balu Krishna S, Hasan Shaikh, Amal Joseph Varghese, Simon Pavamani, and Jeny Rajan, "Automated Segmentation of Head and Neck Cancer from CT Images Using 3D Convolutional Neural Networks"* – Submitted to the *Journal of Imaging Informatics in Medicine (JIIM)*.

Achievements:

- Poster on *"Automated Segmentation of Head and Neck Cancer from CT Images Using 3D Convolutional Neural Networks"* won the **Third Prize** at the **2nd Annual Winter Symposium 2025**.

CHAVI

Enhancing Oncology Imaging Data for AI-driven Research and Public Repository Availability

Joanna Sharon Jagadish[1], Hasan Shaikh[2], Hannah Mary Thomas T[2], Balu Krishna Sasidharan[2], Simon Pavamani[2],

Santam Chakraborty[3], Indranil Mallick[3]

[1] School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India

[2] Department of Radiation Oncology, Christian Medical College (CMC), Vellore, Tamil Nadu, India

[3] Department of Radiation Oncology, Tata Medical Center (TMC), Kolkata, India

Abstract

Medical imaging is a critical pillar of oncology, underpinning cancer diagnosis, treatment planning, and research. However, the development of artificial intelligence (AI) and machine learning (ML) models in oncology requires access to large-scale, high-quality, and standardized imaging datasets. Public repositories such as the Comprehensive Archive of Imaging in Oncology (CHAVI) offer a promising avenue for sharing de-identified imaging datasets. This project focuses on streamlining the integration of Head and Neck Cancer imaging and clinical data into the CHAVI repository by implementing structured data extraction, robust de-identification protocols, and rigorous quality control. Through cleaning, anonymizing, and standardizing datasets, this work enhances data accessibility, supports AI research in oncology, and ensures compliance with international privacy regulations. Ultimately, this project contributes to building a foundation for future AI-driven innovations in early cancer detection, prognosis prediction, and personalized treatment strategies.

1. Introduction

Medical imaging plays a vital role in the diagnosis, treatment, and ongoing management of cancer patients. With the advent of artificial intelligence (AI) and machine learning (ML) technologies, there is a growing demand for large, structured, and high-quality imaging datasets to train and validate predictive models that can enhance clinical decision-making. However, the availability of such datasets remains limited, particularly in the South Asian context, where region-specific imaging biobanks are scarce.

Public repositories such as The Cancer Imaging Archive (TCIA), Genomic Data Commons (GDC), PhysioNet, and the UK Biobank have significantly contributed to the democratization of medical data for research purposes. These platforms enable researchers worldwide to access curated, anonymized datasets under strict ethical and legal guidelines, fostering reproducibility, collaboration, and innovation. Despite these global efforts, many regions still face challenges related to data accessibility, standardization, and ethical sharing.

The Comprehensive Archive of Imaging in Oncology (CHAVI) is an Indian initiative aimed at creating a centralized repository of de-identified oncology imaging data. However, preparing clinical and imaging datasets for integration into CHAVI presents several challenges, including inconsistencies in data formats, patient privacy concerns, and the need for compliance with international regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR).

This project aims to bridge the gap between raw hospital data and research-ready datasets by developing a streamlined process for data preparation and integration. By focusing on Head and Neck Cancer imaging data from Christian Medical College (CMC) Vellore, the project ensures that datasets are cleaned, anonymized, standardized, and ethically compliant, thereby enhancing their utility for AI-driven oncology research and facilitating broader collaborations across the research community.

2. Materials and Methods

2.1 Project Objective

The primary objective of the project was to facilitate the secure, seamless, and efficient integration of Head and Neck Cancer imaging and clinical data into the CHAVI system. Emphasis was placed on designing and implementing a robust data upload pipeline that not only meets CHAVI's technical requirements but also adheres to strict ethical, regulatory, and privacy standards, including HIPAA and GDPR compliance.

A key focus was on implementing comprehensive de-identification protocols. All identifiable patient data, including Protected Health Information (PHI) and Personally Identifiable Information (PII), were systematically anonymized before being uploaded to the repository. This ensured responsible data sharing without compromising patient confidentiality.

2.2 Data Preparation and Cleaning

Initial clinical and imaging datasets often contained inconsistencies, including repeated patient identifiers, missing fields, and non-standardized data formats. To address these issues, an automated preprocessing pipeline was developed that:

- Detected and corrected formatting errors,
- Completed missing fields where possible through logical inference,
- Standardized all data fields according to CHAVI's structured input requirements,
- Removed or anonymized all identifiable patient information.

This process ensured that the datasets were clean, structured, reliable, and ready for integration into CHAVI for future research applications.

2.3 De-identification Protocols

A rigorous de-identification protocol was established to ensure that all sensitive information was removed from clinical and imaging datasets. Steps included:

- Removing patient names, hospital identifiers, and other PII from imaging metadata (e.g., DICOM headers),
- Masking or deleting identifiable free-text fields in clinical reports,
- Assigning unique anonymized identifiers to maintain data integrity across imaging and clinical records without linking back to patient identity.

This ensured full compliance with research ethics guidelines and regulatory frameworks for data protection.

2.4 Data Standardization for AI-readiness

To make the datasets suitable for AI applications, the project emphasized consistent structuring and formatting:

- Clinical data was mapped into standardized formats ensuring interoperability across different research platforms,
- Imaging data was harmonized, ensuring consistency in imaging parameters and metadata fields,
- Datasets were linked appropriately using non-identifiable keys to enable multimodal research (e.g., combining clinical, imaging, and outcome data).

This structured approach laid the groundwork for enabling the development of robust AI models for oncology applications.

2.5 Ethical Compliance and Regulatory Adherence

Throughout the project, all processes were designed to align with internationally accepted standards for data sharing and privacy. Institutional approvals were obtained where necessary, and ongoing documentation ensured that ethical compliance was maintained at every stage of data handling.

Automatic Segmentation

Automated Segmentation of Head and Neck Cancer from CT Images Using 3D Convolutional Neural Networks

Piyush Prabhanjans[1], Aparna V K[1], Rajendra Benny Kuchipudi[2], Hannah Mary Thomas T[2], Balu Krishna S[2], Hasan Shaikh[2], Amal Joseph Varghese[2], Simon Pavamani[2], Jeny Rajan[1]

[1] Department of Computer Science and Engineering, National Institute Karnataka, Surathkal, Mangalore, India

[2] Department of Radiation Oncology, Christian Medical College (CMC), Vellore, Tamil Nadu, India

Abstract

Head and neck cancer (HNC) demands highly precise tumor delineation for effective radiotherapy planning. Manual segmentation is labor-intensive and introduces inter-observer variability, while access to advanced

imaging modalities such as PET/CT remains limited in resource-constrained settings. This study aimed to develop a fully automated, CT-based segmentation model to delineate the gross tumor volume (GTV) of head and neck cancers using a self-configurable 3D deep learning framework, specifically the 3D nnU-Net (version 2). The model incorporates advanced preprocessing techniques and was evaluated using three-fold cross-validation across two datasets: the public HEAD-NECK-RADIOMICS-HN1 dataset and CMC HNC dataset. The proposed model achieved an average Dice Similarity Coefficient (DSC) of 0.72 and a 95th percentile Hausdorff Distance (HD95) of 15.74 mm. These results demonstrate that our CT-only deep learning approach provides a robust, cost-effective solution to improve segmentation accuracy and streamline radiotherapy planning in head and neck cancer care.

1. Introduction

Head and neck cancers present substantial challenges for radiotherapy planning due to their complex anatomy and proximity to critical organs. Accurate segmentation of tumor volumes is essential for delivering high radiation doses to tumors while minimizing exposure to surrounding healthy tissues. Traditionally, tumor delineation has relied on manual segmentation by radiation oncologists, a process that is subjective, time-consuming, and prone to inter-observer variability.

Automated segmentation methods using deep learning, particularly convolutional neural networks (CNNs), have shown promise in enhancing accuracy and efficiency. Although multi-modality imaging (CT, PET, MRI) improves tumor detection, PET and MRI are often not available in low-resource settings. Thus, there is a significant need for reliable CT-based automated segmentation models that are accessible and cost-effective.

In this study, we developed a fully automated tumor segmentation model using the 3D nnU-Net framework, leveraging only CT imaging, to address the need for scalable and reproducible segmentation in head and neck cancer patients, particularly in low- and middle-income countries.

2. Materials and Methods

2.1 Data Acquisition

We utilized two datasets:

- **HEAD-NECK-RADIOMICS-HN1 dataset** from The Cancer Imaging Archive (TCIA) comprising 137 patients (136 after exclusion of one case due to technical issues), and
- **HNC private dataset** from Christian Medical College (CMC), Vellore, consisting of 30 de-identified CT scans.

All patients had histologically confirmed head and neck cancer and were treated with radiotherapy and/or chemotherapy. Gross tumor volumes (GTVs) were manually delineated by experienced Radiation Oncologists and reviewed independently by a second Radiation Oncologist to ensure contouring consistency.

2.2 Data Splitting

Patients were split into training and testing sets in a patient-wise manner to avoid data leakage.

- From HN1: 122 patients for training, 14 for testing
- From HNC: 20 patients for training, 10 for testing

This resulted in a combined training set of 142 patients and a testing set of 24 patients. A three-fold cross-validation strategy was adopted for robust model evaluation.

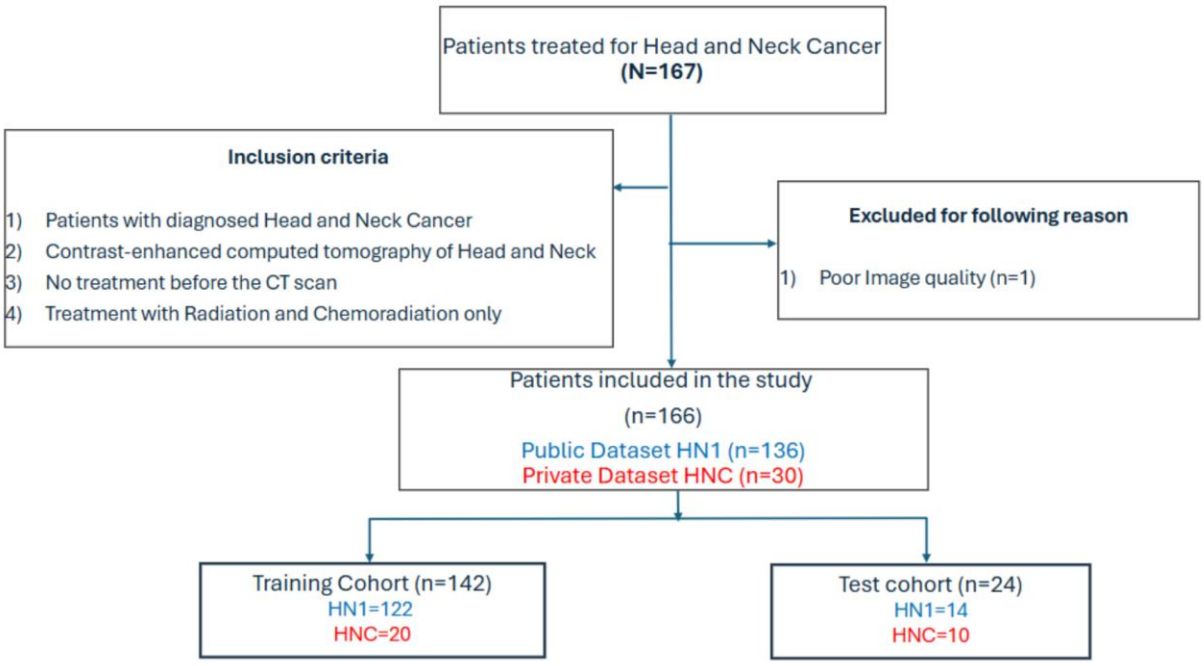

Figure 1: Patient Selection and Dataset Splitting Flowchart

2.3 Data Preprocessing

CT images and corresponding GTV masks were cropped to 256 × 256 pixel regions centered around the head and neck area. Windowing with a Window Level (WL) of 40 and Window Width (WW) of 400 was applied to enhance soft tissue contrast and highlight tumor regions. These standardized preprocessing steps reduced background noise and computational complexity.

2.4 Model Architecture

The 3D nnU-Net (version 2) framework was employed to automatically configure preprocessing, training parameters, and network architecture based on dataset characteristics ("data fingerprint" and "pipeline fingerprint").

The model followed an encoder-decoder structure with multiresolution feature extraction, deep supervision, and automated hyperparameter tuning. Experiments were conducted with residual encoder variations (Medium and Large configurations) to optimize performance. The optimization was performed using Stochastic Gradient Descent (SGD) with Nesterov momentum.

A hybrid loss function combining Dice Loss and Cross-Entropy Loss was used to improve both region-level overlap and pixel-wise classification. Deep supervision was applied at intermediate decoder layers to enhance training stability.

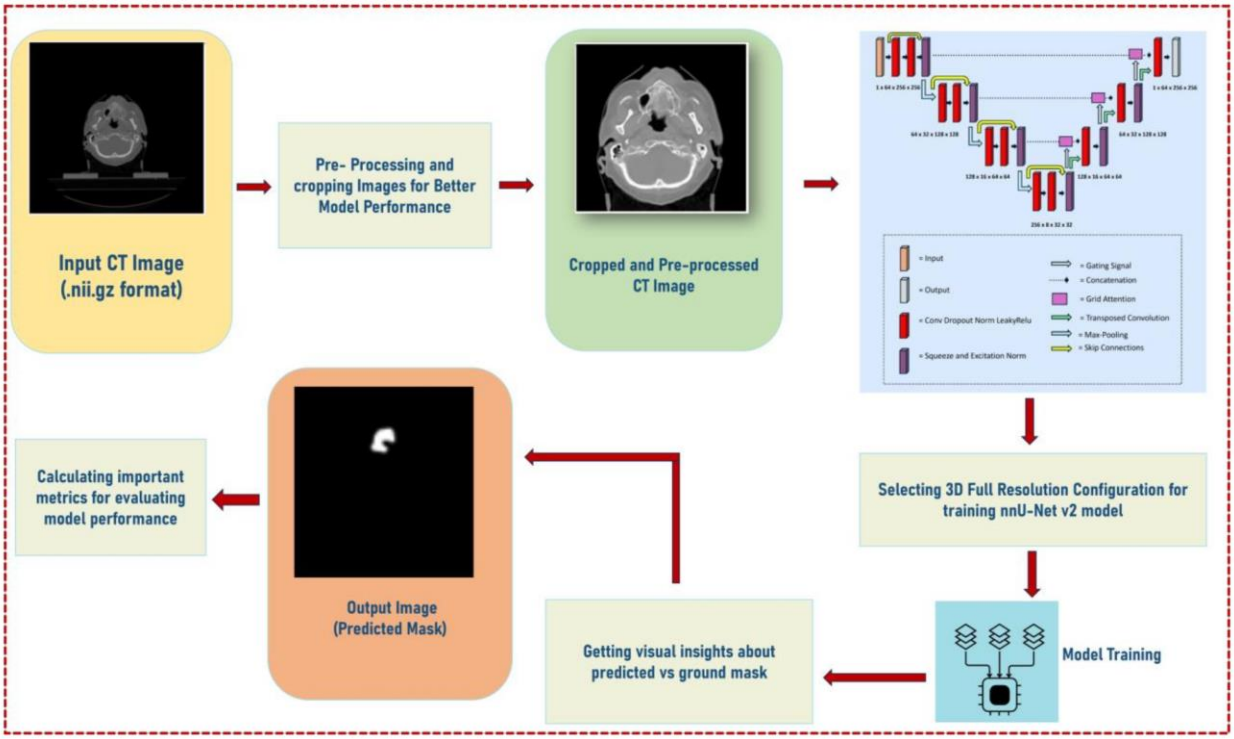

Figure 2: Proposed framework for our work

3. Results and Discussion

The proposed model achieved the following performance across the combined dataset (HN1 + HNC) over three folds of cross-validation:

- Average Dice Similarity Coefficient (DSC): **0.72**
- Average 95th percentile Hausdorff Distance (HD95): **15.74 mm**
- Average Precision: **0.65**
- Average Recall: **0.82**

These results highlight the model's strong segmentation capability using only CT imaging, with reliable tumor boundary delineation and reduced segmentation errors. Comparative evaluations suggested that including additional datasets could further improve generalizability.

TABLE 1: Average Results of Three-Fold Cross Validation Across Combined and Individual Datasets

Dataset	DSC	Precision	Recall	HD95 (in mm)
HN1	0.76	0.71	0.83	12.67
HNC	0.63	0.55	0.80	20.05
HN1 + HNC	0.72	0.65	0.82	15.74

4. Conclusion

This study demonstrates the potential of a CT-only based automated tumor segmentation approach using the 3D nnU-Net framework in head and neck cancer. Achieving an average Dice Similarity Coefficient of 0.72, the model offers a practical, scalable solution for radiotherapy planning, especially in resource-limited settings. Future work will focus on further enhancing performance through bounding box localization (e.g., MedSAM integration), expanding datasets with greater demographic diversity, and employing data augmentation strategies. Such improvements will contribute to more personalized, accurate, and efficient treatment planning in head and neck oncology.

Data Collection Pipeline

Data preparation pipeline for artificial intelligence in Radiation Oncology: Lessons from a large Prospective Imaging trial

Hannah Mary Thomas T[1], Amal Joseph Varghese[1], Manu Mathew[1], Devakumar D[2], Aparna Irodi[3], Joel T[1], Timothy Peace[1], Henry Finlay Godson[1], Rajesh Isiah[1], Simon Pavamani[1], Balu Krishna Sasidharan[1], Hasan Shaikh[1]

[1] Department of Radiation Oncology, Christian Medical College (CMC), Vellore, Tamil Nadu, India

[2] Department of Nuclear Medicine, Christian Medical College (CMC), Vellore, Tamil Nadu, India

[3] Department of Radiology, Christian Medical College (CMC), Vellore, Tamil Nadu, India

Abstract

The integration of Artificial Intelligence (AI) into clinical practice demands robust data preparation pipelines, especially in Radiation Oncology where challenges in data quality and access persist. This work presents a bottom-up approach for AI implementation in Indian Radiation Oncology, focusing on a prospective observational trial for predicting locoregional recurrence in head and neck cancer. The data preparation pipeline encompasses patient consent, clinical data collection, image acquisition, annotation, curation, de-identification, and storage. The workflow, established in 2020, has successfully imaged over 1550 patients. The Data preparation involves metadata analysis, automated quality control, and integration with open-source platforms like Orthanc DICOM server and XNAT. While providing practical insights for data preparation for AI research, the work emphasizes the importance of tailored frameworks for specific healthcare contexts, outlining current practices and future considerations in the evolving landscape of AI in Radiation Oncology.

1. Introduction

The development of AI solutions that are reproducible and transferrable to clinical practice require access to large scale data for model training and optimization. Despite Radiation Oncology departments generating large volumes of imaging data and clinical routinely, quality and access to this data for AI research poses significant challenges in practice.

Implementing Artificial Intelligence (AI) in a clinical environment is a complex task. It requires resources, time, technical capacity, and multidisciplinary perspec-tives. Many AI frameworks outline a set of broad principles and values, but do not provide practical guidance on how these principles can be implemented and trade-offs negotiated. Such enabling factors were yet to be articulat-ed within the context of Radiation Oncology in India, signaling the need for a bottom-up study, and localized models. In this work we report our data preparation pipe-line for AI projects that involve imaging and clinical data. It was aimed to integrate seamlessly within the existing clinical workflow.

This data preparation pipeline was created as part of a prospective observational imaging trial that is on-going with the primary endpoint being Radiomics models for predicting locoregional recurrence in locally advanced head and neck cancer.

2. Materials and Methods

We have subdivided the data preparation pipeline for AI projects into multiple tasks that are typically performed in sequence. Figure 1

Patient consenting: This study was approved by the Institutional Review Board. The patients treated in Radiation Oncology for locally advanced head and neck cancer were explained the observational nature of the trial and consent was obtained for their clinical and imaging data to be used for research purposes.

Clinical Data collection and curation: Clinical data are stored in disparate sources. The demographics and clinical investigations are collected and stored in our electronic medical records and managed in our hospital information system. Radiation Oncology related data such as radiation treatment notes, treatment related toxicity, treatment outcomes etc. are collected and stored in the Aria Oncology Information system. Study specific clinical data are curated, pseudonymized and stored in a Redcap database1.

Image acquisition: Planning CT images were acquired on a Siemens SOMATOM AS. The contrast CT acquisition protocols were standardized for Head and Neck cancer for both machines. This was done after extensive phantom experiments and optimized for field of view, tube current, tube voltage, slice thickness and Image reconstruction kernels ideal for treatment planning and imaging re-search.

Image annotation: The CT images were annotated using EclipseTM Treatment Planning System (Varian Medical Systems, Palo Alto, CA). Templates were created for consistent naming convention using the Eclipse Structure Templates. The gross tumor volume (GTV) was manually delineated by junior radiation oncologists and approved by senior oncologists. All contours selected for the radiomics study were reviewed by radiation oncologists with more than 15 years of experience. Repeat blind segmentation was performed for 20% of the cases to ac-count for inter-observer variability.

Image data curation: The CT and tumour contours were moved from the Oncology Information System to a lightweight, enterprise-grade Orthanc DICOM server. The DICOM server was deployed in a Docker container in a dedicated Research Workstation.

The DICOM metadata of the stored medical images were collected using the REST API of Orthanc. This was automated by creating a bespoke py-thon module to analyze the DICOM metadata, specifically to verify the images for Study Description, Modality Spe-cific image parameters (e.g. CT) and its associated series (DICOM segmentation), perform data quality control, ensure image integrity, query the images, and store the file's orthanc ids in a JSON format. We also use script to retrieve needed data from the Orthanc server using the API, to check for any inconsistently labelled RT structure sets and rename them before further analysis.

Image de-identification: The curated images were pseudo-anonymized by using MIRC Clinical Trial Processor (CTP), an open-source java-based application specifically designed to remove protected health information from DICOM objects2.

All protected health information was replaced with pseu-donyms or unique identifiers and the master table was well-secured and separately stored with access provided only to the researchers involved.

Image storage and Backup: The images and associated DICOM objects from TPS were backed up via S3 bucket on a private cloud setup on the Institutional research cluster and automated using customized python scripts.

The pseudonymized images were sent directly from CTP to XNAT, which is an open-source imaging informatics platform3. XNAT was set up in a Docker container and it allowed data to be stored for separate studies. It also provides programmatic script access to perform AI related workflows using the data from XNAT and route the ana-lyzed data back to the database.

3. Results

Figure 1 shows the data preparation pipeline. The image acquisition, image annotation and basic storage to the research environment was set up in 2020. It was used to prospectively image about 1550 patients till date treated for head and neck cancer in the department of Radiation Oncology. Other components in the workflow have been included and audited asynchronously.

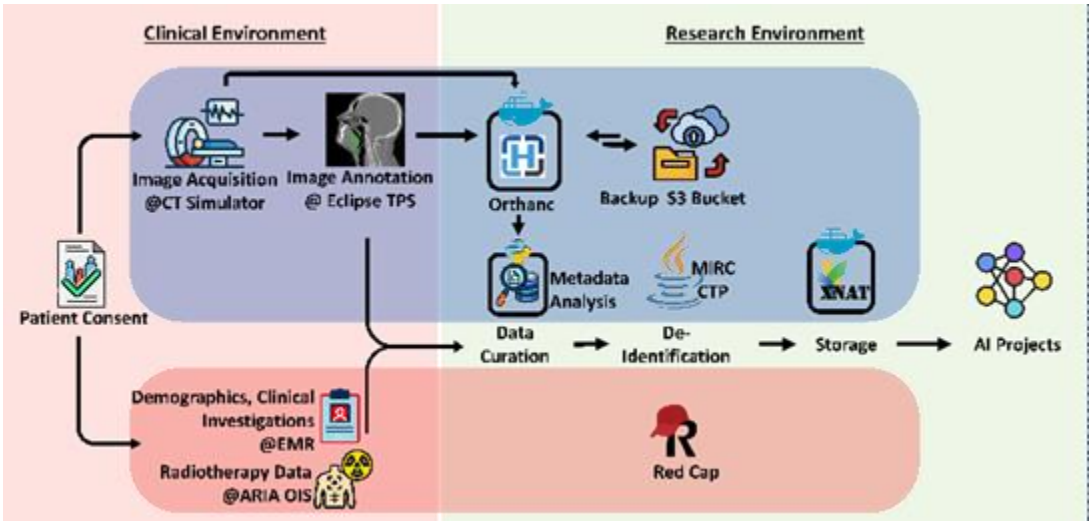

Figure 1: Data preparation pipeline for AI research in Radiation Oncology department

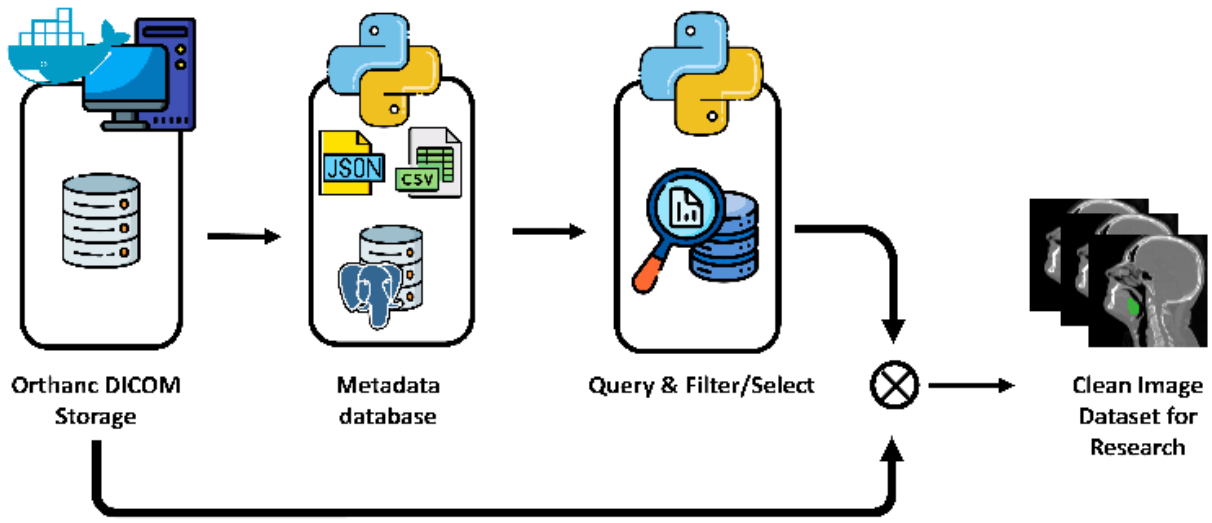

Figure 2: Image data curation workflow

Figure 2 describes the Image Data curation workflow. The Metadata was collected for images received on the Or-thanc Storage. The metadata was analyzed to include only patients related to the study in question. The DICOM metadata tags included for analysis used Personal Health identifier information (Hosp ID, Patient Name etc.) Study (Study Instance UID, Date of study etc.) Series related information (Series description, Modality etc.) SOP In-stance UID unique for each File and the corresponding File Path. After the completion of metadata analysis, the cleaned data was moved forward for de-identification and storage.

4. Discussion

The healthcare sector itself is not a monolith and often necessitates frameworks tailored to specific use cases or applications. In this work we share the data preparation pipeline that allows us to integrate clinical and research environments to get the data ready for AI projects.

The choice to use open-source platforms in the research environment was to ensure those processes are translata-ble to other institutions/groups when needed. 3D Slicer was initially set up as our DICOM listener for image data curation. However, although 3D Slicer offers many capa-bilities, it was not a lightweight solution and could not handle receiving large volumes of data. We migrated to Orthanc DICOM server for receiving the data from the TPS for the as it seamlessly receives and stores data whilst large volume of data transfer happens from our multiple imaging data sources (e.g. CT simulator, TPS, PACS) for different studies.

The range of DICOM metadata tags available for querying and/or retrieval from any DICOM based PACS is limited and designed for everyday operational use rather than for audit or research purposes. As with other reported studies 4–6 we found query retrieval to be simpler on the DICOM metadata when it is stored independently from the image. We created a bespoke python-based module that could be selectively used to ensure data integrity (e.g.. ensuring availability of RTSTRUCT related to an imaging series). This allows us to check for missing data or audit the imaging protocol. This module can also be utilized as an automated procedure for all patients within the study of choice prior to visual inspection.

As reported by other research groups 7,8, XNAT was chosen for our research data since the access is built on a REST architecture which offers flexibility while interact-ing with the data. For example, it allows us to integrate the radiomics feature extraction on the data stored and conveniently stores the radiomics features back into the database for each individual image set. It also provides the flexibility to introduce federated learning-based AI models to learn on the data.

There are some limitations and challenges that still need to be resolved. For e.g. the data entry for clinical data is still manual entry which is both time-consuming and error prone. Next steps include creation of an API that could capture some of the data fields from the EMR directly to the Redcap database.

5. Conclusion

Our work describes a meticulous data preparation pipeline tailored for AI integration in Radiation Oncology. This work emphasizes the importance of context-specific frameworks in addressing clinical challenges. The presented pipeline contributes to the discourse on AI implementation in Radiation Oncology, providing insights for future advancements in the field.

References

1. REDCap. https://www.project-redcap.org/
2. MIRC CTP - MircWiki. https://mircwiki.rsna.org/index.php?title=MIRC_CTP
3. XNAT - Home. https://www.xnat.org/
4. Santos, M. & Rocha, N. P. Outcomes from Indexing Initiatives of Medical Imaging DICOM Metadata Repositories. A Secondary Analysis. Procedia Comput. Sci. 138, 203–208 (2018).
5. Mackenzie, A., Lewis, E. & Loveland, J. Successes and challenges in extracting information from DICOM image databases for audit and research. Br. J. Radiol. 96, 20230104 (2023).
6. Kathiravelu, P. et al. A DICOM Framework for Machine Learning and Processing Pipelines Against Real-time Radiology Images. J. Digit. Imaging 34, 1005–1013 (2021).
7. Gutman, D. A. et al. Web based tools for visualizing imaging data and development of XNATView, a zero footprint image viewer. Front. Neuroinformatics 8, 53 (2014).
8. XNAT Soup, a data finding utility used to visualize col | Open-i. https://openi.nlm.nih.gov/detailedresult?img=PMC4034701_fninf-08-00053-g0007&query=&req=4
9. Kundu S, Chakraborty S, Chatterjee S, Das S, Achari RB, Mukhopadhyay J, et al. De-Identification of Radiomics Data Retaining Longitudinal Temporal Information. J Med Syst. 2020 Apr 2;44(5):99.
10. Kundu S, Chakraborty S, Mukhopadhyay J, Das S, Chatterjee S, Basu Achari R, et al. Research Goal-Driven Data Model and Harmonization for De-Identifying Patient Data in Radiomics. J Digit Imaging [Internet]. 2021 Jul 9; Available from: https://doi.org/10.1007/s10278-021-00476-9
11. Kundu S, Chakraborty S, Mukhopadhyay J, Das S, Chatterjee S, Achari RB, et al. Design and Development of a Medical Image Databank for Assisting Studies in Radiomics. J Digit Imaging [Internet]. 2022 Feb 15; Available from: http://dx.doi.org/10.1007/s10278-021-00576-6