4. Image Acquisition[2]

Medical image biomarkers of cancer have improved patient care through customized healthcare practices [163]. Biological imagery features are more advantageous than genomic biomarkers. A non-invasive procedure can characterize a heterogeneous tumor effectively [164]. On the other hand, in the invasive process, only limited tissue remains available for biopsy. The Cancer Genome Atlas (TCGA) data is mainly focused on mapping cancer phenotypes to genotypes. It provides clinical images matched to subjects from The Cancer Genome Atlas (TCGA). The Genomic Data Commons (GDC) data portal stores the clinical data while The Cancer Imaging Archive (TCIA) stores the radiological data [12]. In TCGA, case identifiers allow researchers to explore databases for finding correlated genotype, phenotype, and treatment outcomes. Tissues for TCGA are collected from different sites across the world. The heterogeneous image data sets vary in terms of scanner modalities, and acquisition protocols. In most cases, the images are acquired while carrying out routine care. It does not follow a controlled research study. This makes the dataset independent of any predetermined bias or correlation. A significant contribution from different institutions to the archive creates multi-institutional data sets and they become an open resource for researchers. As per the methodology adopted by TCGA, the comprehensive quantification of tumor phenotypes are done by applying a large number of quantitative image features. Features quantifying tumor image intensity, shape, and texture are extracted. A large number of radiomic features have been re-invented as having significant prognostic power in these independent data sets. Analysis has revealed that a predictive radiomic signature is linked with underlying gene-expression patterns. These suggest that radiomics identifies a general prognostic phenotype in various types of cancers. A significant scientific impact may be observed as imaging is normally used in clinical practice. This provides an unprecedented opportunity to improve the decision-support system which may help in cheaper cancer treatment.

In this research work, eight types of cancers have been included for classification: Urothelial Bladder Carcinoma, Cervical renal papillary cell carcinoma, Liver Hepatocellular Carcinoma, Non-Small Cell Lung Cancer, Uterine Corpus Endometrial

[2] Based on author's publication no. 9 [Appendix D]

Carcinoma, Thyroid Cancer, head and neck squamous cell carcinoma (HNSCC), and Breast Invasive Carcinoma. Thus, eight different datasets from The Cancer Imaging Archive (TCIA) have been considered for malignant tumor classification. Each of these datasets represents one particular type of tumor based on its originating organ. In this way, eight different originating organs have been covered: prostate, kidney, liver, lung, uterus, thyroid glands, head & neck, and breast.

4.1 TCGA-BLCA

TCGA-BLCA dataset represents Urothelial Bladder Carcinoma (BLCA) which is the most common type of bladder cancer and one of the leading causes of cancer-related death worldwide. The dataset was created by Ken Clark, last modified by Quasar Jarosz on Mar 30, 2020 [111]. The dataset comprises 111,781 images of 120 numbers of patients. Major imaging modalities are Computed Tomography (CT), Magnetic Resonance (MR), Computed Radiography (CR), Positron Emission Tomography (PET), and Digital Radiography (DX).

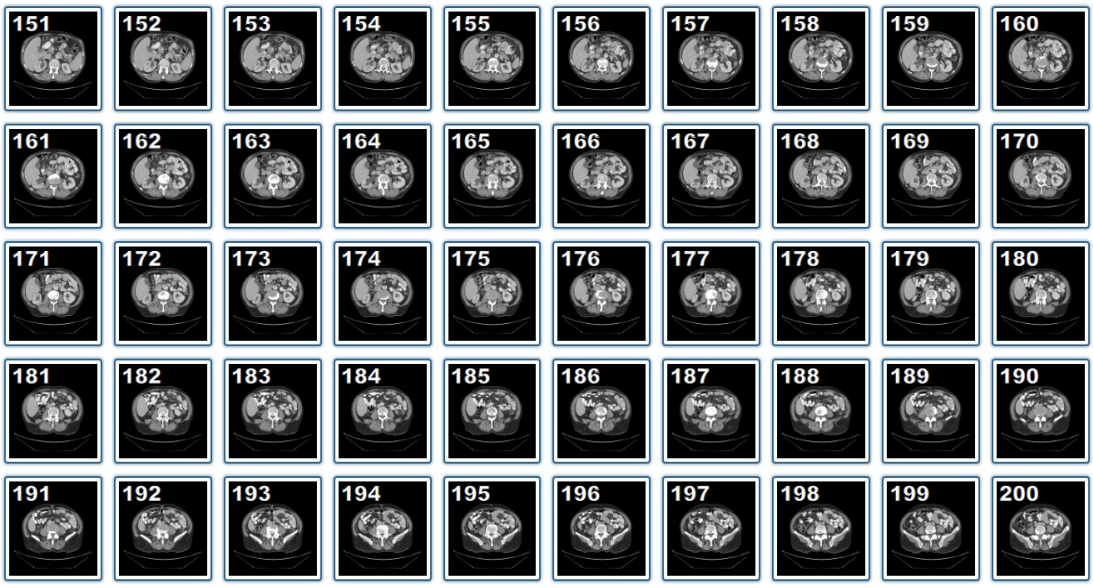

Figure 16: Glimpse of the image set from TCGA-BLCA (patient barcode: TCGA-2F-A9KO)

The following institutions have contributed towards the database development:

- University of North Carolina
- Barretos Cancer Hospital, Brazil
- University of Chicago

- University of Sheffield
- Memorial Sloan-Kettering Cancer Center, New York
- Lahey Hospital & Medical Center, Burlington
- University of Southern California

4.2 TCGA-KIRP

TCGA-KIRP depicts cervical renal papillary cell carcinoma which is one of the most common kidney-related cancers. It was created by Justin Kirby and was last modified by Quasar Jarosz on Mar 30, 2020 [112]. It has 33 cases consisting of 376 series and 26,667 images. Major imaging modalities are CT, MR, and PT.

Figure 17: Glimpse of TCGA-KIRP image dataset (patient barcode: TCGA-B9-4113)

The following individuals have provided data for this collection:

- National Cancer Institute, Bethesda
- University of North Carolina
- Roswell Park Cancer Institute, New York
- Lahey Hospital & Medical Center, Burlington

4.3 TCGA-LIHC

TCGA-LIHC is the Liver Hepatocellular Carcinoma (LIHC) image dataset. It was created by Justin Kirby and was last modified by Tracy Nolan on Jan 08, 2020 [113]. 97 cases have been incorporated here. The total number of series is 1688 having a total number of 125,397 images. Major imaging modalities are CT, MR, and PT.

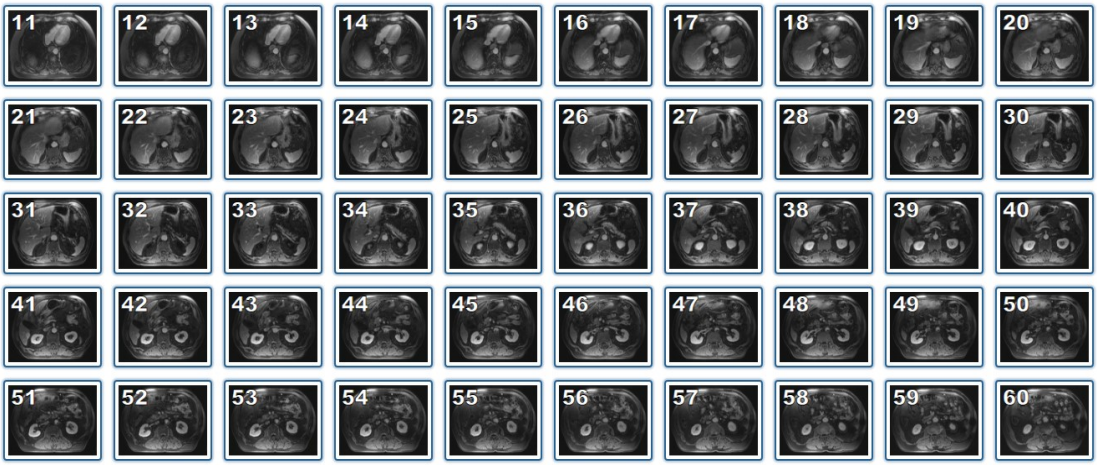

Figure 18: Glimpse of TCGA-LIHC Images (patient barcode: TCGA-BC-4073)

The following institutions have contributed data for this collection:

- Mayo Clinic, Rochester, MN
- The University of North Carolina,
- Alberta Health Services
- Lahey Hospital & Medical Center, Burlington

4.4 NSCLC Radiogenomics

It is a well-known radiogenomic dataset having 211 subjects of Non-Small Cell Lung Cancer (NSCLC) [114]. It was created by Kirk Smith and was last modified by Justin Kirby on Apr 01, 2020. The dataset comprises different imaging modalities like Computed Tomography (CT), Positron Emission Tomography (PET)/CT images. It also contains semantic observations and segmentation maps of tumors. Quantitative imaging data are mapped with clinical data, including pathological stage or grade of tumor, mean or maximum uptake values (in case of PET), treatment strategy, and survival outcomes [115]. It also depicts the demographic details of patients such as gender, age, smoking habits, location, etc. This dataset also helps in finding the inherent bond between genomic and medical image attributes. In this way, it facilitates the development of predictive biological features of medical images. Eminent doctors from Stanford University have constituted the database.

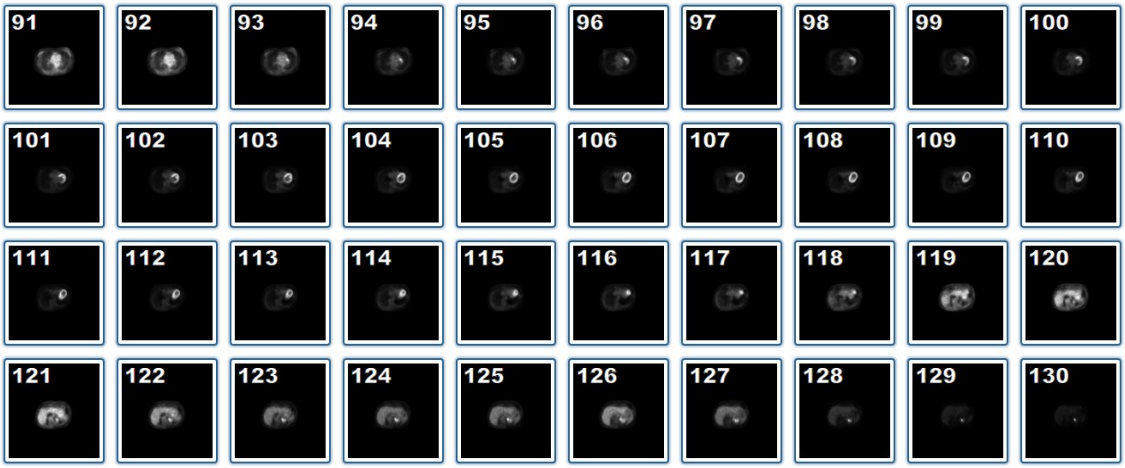

Figure 19: TCIA NSCLC Radiogenomics (Case ID R01-001)

4.5 TCGA-THCA

TCGA-THCA represents thyroid cancer which is a common sub-type of head and neck cancer. It was created by Ken Clark and was last modified by Quasar Jarosz on Mar 30, 2020. There are 6 cases in the image set with 28 series and 2,780 numbers of images [161]. Major imaging modalities are CT and PET.

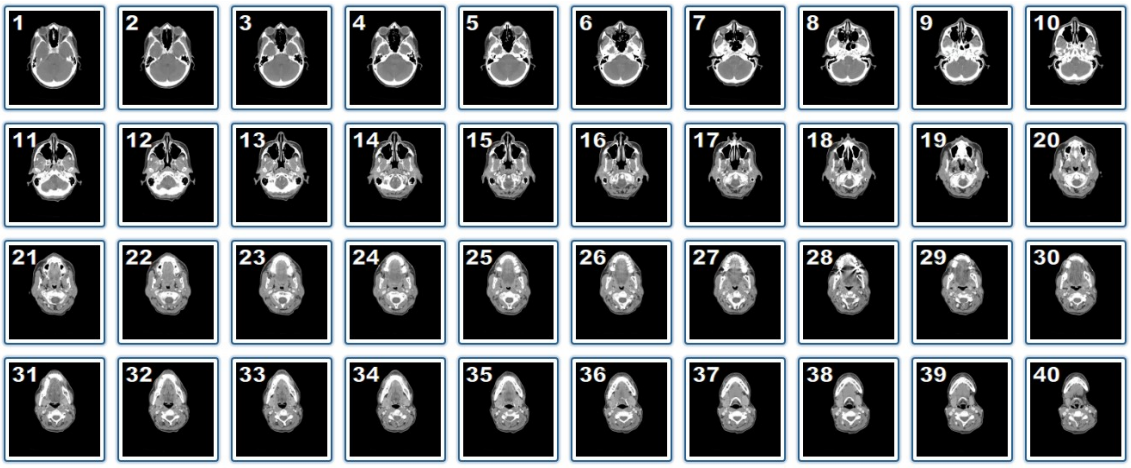

Figure 20: Glimpse of TCGA-THCA (patient barcode: TCGA-DE-A4MA)

The following institutions/individuals have provided data for this collection:

- The University of North Carolina,
- Roswell Park Cancer Institute, New York

4.6 TCGA-UCEC

TCGA-UCEC represents the Uterine Corpus Endometrial Carcinoma. It was created by Ken Clark and was last modified by Quasar Jarosz on Mar 30, 2020 [162]. There are 65 cases including 912 series having a total of 75,829 numbers of images. Major imaging modalities are CT, CR, MR, and PT.

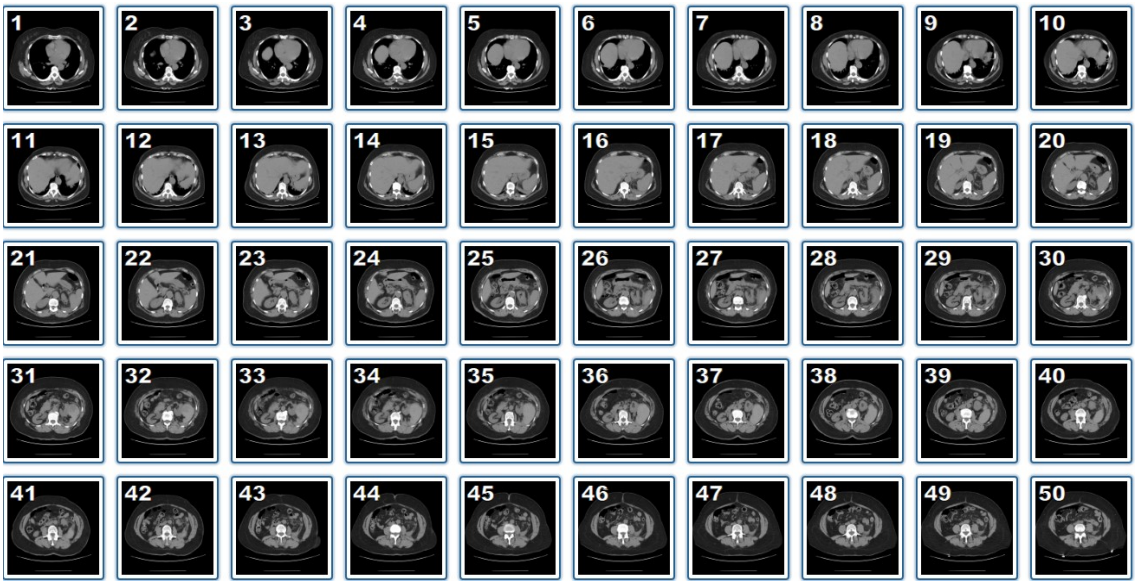

Figure 21: Glimpse of TCGA-UCEC (patient barcode: TCGA-D1-A0ZO)

The following institutions/individuals have provided data for this collection:

- Mayo Clinic, Rochester
- Washington University
- MD Anderson Cancer Center, Houston

4.7 Head-Neck-Radiomics HN1

Radiomics is the complete numerical representation of the tumor's observed characteristics by applying a large number of quantitative imagery attributes [116]. Clinical data of 137 head and neck squamous cell carcinoma (HNSCC) patients have been clubbed together to form the Head-Neck radiomics collection [117]. These patients were treated by radiotherapy. The main imaging modality is computed tomography (CT). It was created by Geri Blake and was last modified by Michelle Tacconelli on Apr 01, 2020. This dataset is provided as an open repository to support the research in radiomics.

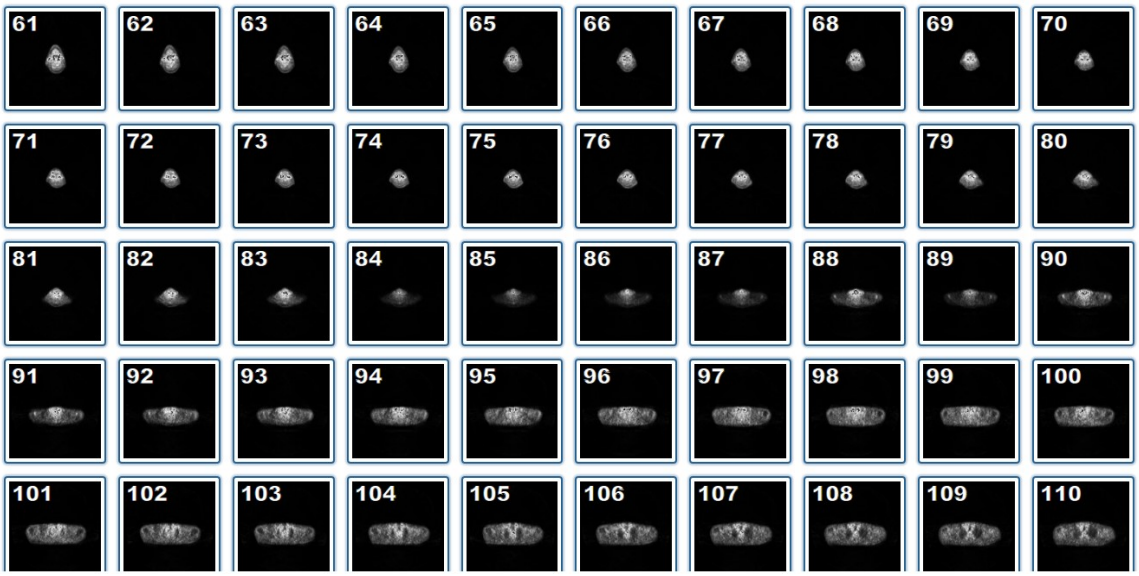

Figure 22: Glimpse of Head and Neck HN1 radiomics (Case ID: HN1004)

The following institutions have provided data for this collection:

- Maastricht University Medical Centre, the Netherlands
- Dana-Farber Cancer Institute, USA
- Harvard Medical School, USA.

4.8 TCGA-BRCA

TCGA-BRCA represents Breast Invasive Carcinoma. The dataset was created by Justin Kirby and was last modified by Quasar Jarosz on Mar 30, 2020 [118]. It has 164 cases with 1877 series containing a total of 230, 167 images. Imaging modalities are MR and mammography (MG).

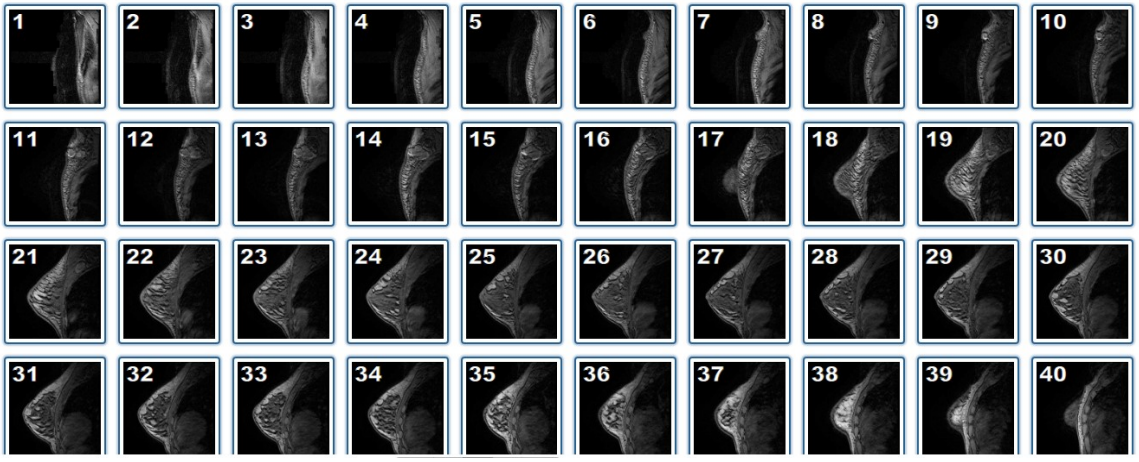

Figure 23: Glimpse of TCGA-BRCA (patient barcode: TCGA-EW-A1PC)

The following institutions have provided data for this collection:

- Mayo Clinic, Rochester
- University of Pittsburgh
- University of Miami
- Memorial Sloan-Kettering Cancer Center, New York
- University of North Carolina
- University of Chicago, Chicago
- Roswell Park Cancer Institute, New York

4.9. Final Image Collection

The final image acquisition has been carried out by retrieving images from all the aforementioned collections. Each subject has a large set of pre-surgical DICOM images stored in the respective database. Each case is identified with a unique Patient ID. Each patient's demographic, clinical, pathological, and/or genomic data are stored in TCGA. In the present research work, the most important features are the TNM staging, AJCC staging, and histological grading embedded in the clinical data against the respective patient. These features have been used as the class or target variable while carrying out the classification task. In some cases, the histological grades or the pathological stage are missing. Those cases are discarded during the final database preparation. Thirty best scans from each case having pathological data from all the eight collections have been taken to form the final image collection. In this way, 717 cases where the supportive clinical and pathological data are present have been considered and from each such case, thirty best scans were extracted. Thus, 21,510 radiological images have been collected to form the new image dataset. This newly prepared image collection is heterogeneous in terms of imaging modality, cancer type, cancer stage/grade, and patients' demographic characteristics.

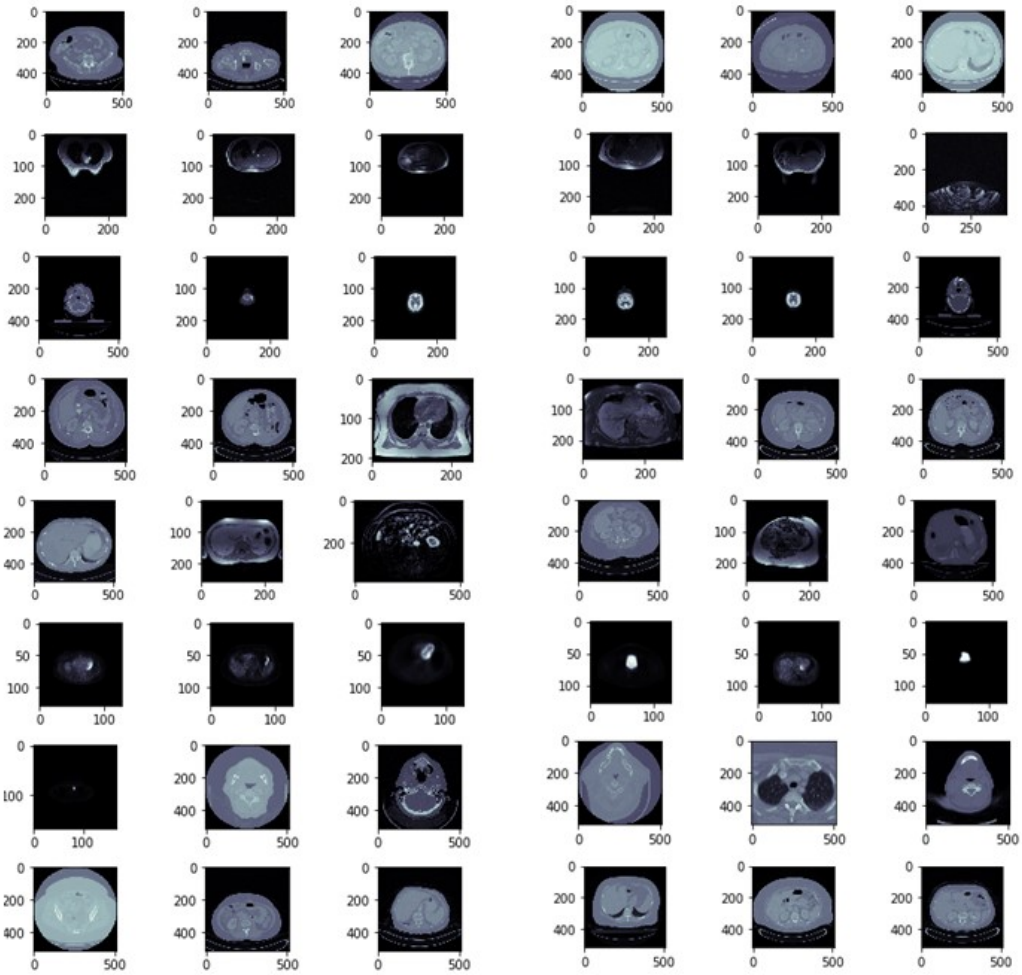

Figure 24: Glimpse of the final image collection (1st row represents TCGA-BRCA, 2nd row represents Head & Neck Radiomics, 3rd row represents TCGA-BLCA, 4th row represents TCGA-KIRP, 5th row represents TCGA-LIHC, 6th row represents NSCLC radiogenomics, 7th row represents TCGA-THCA and 8th row represents TCGA-UCEC)

4.10 Conclusion

In this chapter, the process of developing an image collection of the heterogeneous source has been described. Images from eight different open databases have been collected. Each database represents one particular type of tumor depending upon its originating organ. The cases where clinical information regarding tumor stage and grade are available to have been considered for the final image database preparation. Twenty best scans have been retrieved from each prospective case and clubbed together as a single image collection where each image has a case ID. This has made the new image repository a diversified mix of modalities containing details of patients of different demography. This has indeed brought novelty in the present research work.