Data Science Practicum Project Proposal

*Project Sponsor: Professor Jay Hou*

*Student:* [*Arpit Dang*](mailto:arpit_dang@brown.edu)*,* [*Maxine Baghdadi*](mailto:maxine_baghdadi@brown.edu)*,* [*Xiaoke Song*](mailto:xiaoke_song@brown.edu)*,* [*Yitong W*](mailto:yitong_wang@brown.edu)

# Context

Glioblastoma is the most common form of brain cancer in adults and remains one of the deadliest of human cancers (Stupp et al. 2009; Cai and Sughrue 2018). Survival beyond two years is rare, and therapeutic options limited. Conventional structural magnetic resonance imaging (MRI) is the standard of care imaging method for neuro-oncologic practice. However, the fundamental understanding of tumor progression and patient survival as a function of MRI data remains unclear.

This project aims to bridge that gap by developing a machine learning pipeline to predict cancer stage from multimodal MRI scans and clinical metadata. Using the TCGA-GBM dataset from The Cancer Imaging Archive (TCIA), we will extract tumor-specific features (e.g., volume, shape descriptors) from DICOM-format MRIs and combine them with patient-level clinical variables (e.g., age, sex, molecular biomarkers) to predict WHO CNS cancer grade.

# Goals and OKRs

1. **Objective**: Acquire and process the MRI data from TCGA database
   1. **Timeline:** Between Week 1 (June 18th) to Week 5 (July 16th)
   2. Successfully extract and organize patient DICOM scans and metadata for at least 200 patients
   3. Identify key MRI series and document imaging structure across patients
   4. Implement a scalable preprocessing pipeline that extracts tumor-relevant image slices and preprocesses metadata and outputs a structured dataset ready for feature extraction
2. **Objective:** Build and evaluate a cancer staging prediction model
   1. **Timeline:** Between Week 5 (July 16th) to Week 10 (August 20th)
   2. Extract tumor features (volume, area, etc) from at least 80% of usable patient scans
   3. Link extracted features with clinical data and create a clean modeling dataset
   4. Train and evaluate at least two machine learning classifiers (e.g. random forest, ordinal regression) to predict WHO CNS Grade (1–4)
   5. Achieve > 70% classification accuracy on a held-out test set and report AUC and F1-score for each model

## Data Science Competency Areas

1. Data Wrangling: Extract and normalize MRI volumes using tools like nibabel, parse segmentation masks, and merge data with clinical metadata. This ensures we have a clean and structured dataset for analysis.
2. Exploratory Data Analysis (EDA): Visually and statistically explore relationships between tumor features, clinical metadata, and cancer stage to inform modeling choices.
3. Feature Engineering: Quantitatively describe tumor regions from segmentation masks by computing spatial and morphological features (e.g., volume, centroid, bounding box dimensions, compactness). These features form the foundation of our predictive model.
4. Supervised Learning: Apply and compare classical machine learning models such as random forests, logistic regression, and ordinal regression to predict cancer stage based on engineered features and patient metadata.
5. Model Evaluation: Use metrics such as accuracy, ROC AUC, F1-score, and confusion matrix to rigorously assess classification performance. Employ cross-validation where appropriate.
6. Collaboration and Versioning: Use tools such as Colab, GitHub, and Drive to manage team contributions, maintain reproducibility, and track progress.

Data science is key to this project because it helps us turn complex MRI scans and clinical data into meaningful, structured features we can analyze. It lets us extract things like tumor volume from segmentation masks and combine that with patient info to build models that predict cancer stage, something that wouldn’t be possible just by looking at the raw images.

## Deliverables

1. Pipeline for extracting tumor features from pre-existing segmentation masks
2. Merged dataset combining image-derived features with clinical metadata
3. Machine learning model to predict WHO CNS Grade from these features
4. Final report and presentation outlining methodology, results, limitations and future directions

## Non-Goals

1. We will not train/evaluate a 3D segmentation model (e.g. 3D U-Net or TransUNet)
2. We will not attempt to predict precise survival time

These are excluded due to scope and time limitations.

# Methodology

### Methodology

1. Data Acquisition & Exploration
   1. Download DICOM-format MRI scans and corresponding clinical data from the TCGA-GBM collection on TCIA. Organize and explore the structure of the imaging and metadata files for each patient.
2. Tumor Region Identification
   1. Use DICOM series (e.g., T1-weighted post-contrast) to isolate tumor-relevant slices. If segmentation is not provided, explore options for semi-automated annotation or intensity-based region detection.
3. Feature Extraction from DICOM Images
   1. Extract tumor-related features such as slice-wise tumor area, estimated volume, intensity statistics, and shape descriptors directly from pixel arrays. Merge with patient-level clinical features (e.g., age, IDH status, outcome).
4. Predictive Modeling
   1. Train machine learning models (e.g., random forest, logistic/ordinal regression) to predict cancer stage or a severity proxy using the extracted image features and clinical metadata. Evaluate using classification metrics like accuracy, F1-score, and AUC.

### Data

1. **Source:** TCGA-GBM collection via The Cancer Imaging Archive (TCIA)
2. **Format:** DICOM
3. **Sample Size:** ~262 glioblastoma patients
4. **Components:** Multimodal MRIs + clinical metadata (e.g., age, biomarkers, survival)

# Group Collaboration

The team will meet weekly via Google Meet to discuss progress, divide responsibilities, and resolve blockers. In addition, we will meet weekly with Professor Hou for feedback. Team communication will take place over Slack, with shared files and code maintained in Google Drive and GitHub. Tasks will be split by core competencies (e.g., preprocessing, feature extraction, modeling, evaluation).

# **Project Steps**

| **Task** | **Assigned** | **Notes** | **Completion** |
| --- | --- | --- | --- |
| Set Up Oscar | All | Request access [here](https://brown.co1.qualtrics.com/jfe/form/SV_0GtBE8kWJpmeG4B) |  |
| Import Data | All |  |  |
| EDA |  |  |  |
| Preprocessor |  |  |  |
| Create Models |  |  |  |
| Evaluate Models |  |  |  |
| Create Poster |  |  |  |

# Appendix

### Glossary

| Term | Definition |
| --- | --- |
| DICOM | A standard format for storing and transmitting medical imaging data |
| TCGA | The Cancer Genome Atlas - a large genomic and imaging database |
| WHO CNS Grade | Tumor grading system used to classify the severity of brain tumors |

### Related Prior Work

Chen et al., “Deep Learning-Based Tumor Segmentation on UCSF-PDGM MRI Scans” (2023): Implemented 3D U-Net and TransUNet on the same dataset to predict masks which we use directly in our project.

### Questions

N/A

### Risks and Uncertainties

1. **Class Imbalance:** WHO CNS grades may be unevenly distributed, which can affect classifier performance.
2. **Incomplete Data:** Some subjects may have missing imaging series or incomplete metadata.
3. **Interpretability:** While models may perform well, predictions are not ready for clinical application and must be interpreted cautiously.