EVALUATION OF RADIOMICS AND MACHINE LEARNING IN IDENTIFICATION OF AGGRESSIVE TUMOR FEATURES IN RENAL CELL CARCINOMA (RCC)

by

Sidharth Gurbani

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

(Electrical and Computer Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2020

I dedicate my thesis to my pet dog, Brocky, who sadly passed away very recently.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Professor Varun Jog, for introducing me to the fascinating world of machine learning in medical imaging and for giving me the opportunity and freedom to explore as much of it as possible. His encouragement and scientific supervision helped me to move forward for further success. Special thanks go to Prof. Dane Morgan, Prof. Meghan Lubner and Mingren Shen for their endless amount of patience, support, insight and guidance. This thesis would not have been possible without all of their advice. I would also like to thank my parents and my brother for their never-ending support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

NOMENCLATURE

NG	Nuclear Grade
sarc	Sarcomatoid
noncon	Non Contrast
pv	Portal Venous
MDCT	Multi Detector Computed Tomography
RCC	Renal Cell Carcinoma
CT	Computed Tomography
GLCM	Gray Level Co-occurence Matrix
t-SNE	t-distributed stochastic neighbor embedding
XGB	XG Boost
RF	Random Forest
SVM	Support Vector Machine
CV	Cross Validation
SMOTE	Synthetic Minority Oversampling Technique
AUC	Area Under Receiver Operating Characteristic Curve

EVALUATION OF RADIOMICS AND MACHINE LEARNING IN IDENTIFICATION OF AGGRESSIVE TUMOR FEATURES IN RENAL CELL CARCINOMA (RCC)

Sidharth Gurbani

Under the supervision of Assistant Professor Varun Jog

At the University of Wisconsin-Madison

Varun Jog

12/17/2020

ABSTRACT

The purpose of this study was to evaluate the use of CT radiomics features and machine learning analysis to identify aggressive tumor features, including high nuclear grade (NG) and sarcomatoid (sarc) features, in large Renal Cell Carcinomas (RCCs).

CT-based, volumetric radiomics analysis was performed on non-contrast (NC) and portal-venous (PV) phase multidetector computed tomography (MDCT) images of large (>7 cm) untreated RCCs in 141 patients (46W/95M, mean age 60 years). Machine learning analysis was applied to the extracted radiomics data to evaluate for association with high NG (grade 3-4), with multichannel analysis for NG performed in a subset of patients (n=80). A similar analysis was performed in a sarcomatoid rich cohort (n=43, 31M/12F, mean age 63.7 yrs) using size matched non-sarcomatoid controls (n=49) for identification of sarcomatoid change.

The XG Boost Model performed best on the tested data. After manual and machine feature extraction, models consisted of 3, 7, 5, 10 radiomics features for NC sarc, PV sarc, NC NG and PV NG respectively. The area under the receiver operating characteristic curve (AUC) for these models was 0.59, 0.65, 0.69 and 0.58 respectively. The multichannel NG model extracted 6 radiomic features using the feature selection strategy and showed an AUC of 0.67. We found statistically significant but weak associations between aggressive tumor features (high nuclear grade, sarcomatoid features) in large RCC using 3D radiomics and machine learning analysis.

Chapter 1

Introduction

As the volume of computed tomography (CT) performed for a variety of indications continues to increase, the incidence of renal cell carcinoma (RCC) has also continued to rise [18, 4, 14, 6, 12, 19, 21]. Spatial heterogeneity is a common feature of RCC, with multiple studies demonstrating variability within tumors with respect to pathologic features, genomics, and RNA/protein expression [7, 2, 9]. This heterogeneity gives rise to a spectrum of biologic and clinical behaviors, with an increasingly less aggressive management approach in more indolent disease and nephron sparing approaches in cases where intervention is warranted [25]. Pathologic markers of tumor aggressiveness such as higher nuclear grade (NG) or presence of sarcomatoid (sarc) features may only be present in a small portion of the tumor but may profoundly impact treatment decisions and prognosis. These small areas can be challenging to identify on biopsy, and although radiomic features provide more global tumor assessment and have shown some promise in non-invasively capturing and characterizing tumor heterogeneity, some aggressive tumor features have remained elusive at imaging. If aggressive features could be reliably identified in advance of surgery, either through more targeted biopsies or non-invasive assessment, it could have immediate clinical impact on treatment decisions and prognostication. Recently, multiple groups have used machine learning analysis applied to radiomics features in an attempt to improve performance in identification of aggressive features such as high nuclear grade on imaging, with some success [3, 11, 23, 24, 8, 5, 13]. Identification of sarcomatoid features has remained challenging from CT imaging. The purpose of this study is to evaluate the use of CT radiomics features and machine learning analysis to identify aggressive tumor features, including high nuclear grade and sarcomatoid features, in large

RCCs. For nuclear grade, this would be an attempt to reproduce other groups' results and for sarcomatoid features, to identify an as yet unidentified radiomics imaging signature.

Chapter 2

Methods

This study was IRB approved and HIPAA compliant.

2.1 Patient selection and CT Images

The CT images obtained between 2000 and 2013 of 141 patients (46 women and 95 men, mean age 60 years) with large (>7 cm) RCCs were obtained from the surgical database of the Department of Urology and were retrospectively reviewed. All patients in the cohort had a CT scan performed before undergoing surgery or receiving any other treatment. Subsequent removal of the primary tumor and pathologic analysis that included histologic subtyping and nuclear grading were performed for all patients. CT texture analysis data from these patients were previously analyzed in Lubner et al. 2016 [17] and a multi-platform radiomics analysis was performed by Dreyfuss et al in 2019 [15]. We specifically targeted large RCCs to increase the likelihood that aggressive features would be present.

Analysis of both portal venous phase images (n=124, 44 with portal venous phase only) and non contrast images (n=97; 17 had non contrast images only, 80 had both non contrast and portal venous phase images) were performed. 74 of 124 portal venous (59.7%) CT examinations were performed at institutions other than the study institution. All scans were performed using MDCT scanners and the imaging parameters were as follows: a tube potential of 100-140 kV (with 110 of 124 (89.4%) scans using a tube potential of 120 kV) and a matrix of 512 x 512 x 16. Most CT scans were performed using automated or variable tube current, and the slice thickness used for 122 of 125 scans was 2-5 mm. Although the non contrast and portal

venous analyses were performed separately, a multi-channel analysis of patients who had both datasets was additionally performed. This cohort of 141 patients with large RCCs was used in the assessment of imaging features of nuclear grade. Patients with nuclear grade of 3-4 were considered high grade, while grades 1-2 were considered low grade.

A second sarcomatoid rich dataset was created, using CT imaging obtained between 2001-2018, including 43 RCCs with sarcomatoid features (31M, 12F, mean age 63.7 yrs) with 49 size matched non-sarcomatoid RCCs from the nuclear grade cohort above to serve as controls (30M, 19F, mean age 64.4 yrs) with extracted radiomics features on a similar CT analysis as described above was performed. As with the nuclear grade analysis, both non contrast (n=28, n=3 non contrast only) and portal venous phase CT (n=40, n=15 pv only, n=25 both pv and non contrast) images in patients with sarcomatoid RCCs were evaluated. Size-matched non sarcomatoid controls came from the large RCC nuclear grade dataset and had similar distribution of non contrast and portal venous exams (pv n=49, non con=36, both n=31). For a subset of 25 patients in the sarcomatoid cohort, the percentage of sarcomatoid features present in the tumor was quantified by the surgical pathologist

2.2 Radiomics Platform

Radiomics features were extracted using Healthmyne (Madison, WI, USA), a server-based platform that performs volumetric CT radiomic analysis. Healthmyne does not perform a filtration step and analyzes unfiltered (SSF=0) data. This software extracts over 300 radiomics features, including first-order texture features (mean gray-level intensity, entropy, standard deviation) and second-order texture features derived using gray-level co-occurrence matrix (GLCM). Second-order metrics allow quantification of the spatial relationship between pixels [16]. It also extracts a variety of anatomic and morphologic tumor descriptors including tumor volume, surface area, sphericity etc. Some features are locations in the image used to calculate distances (long axis, short axis etc). These do not extract meaningful image data, only reflect coordinates and were manually excluded from the analysis (the calculated distances from these coordinates reflecting measurements were included).

2.3 Region of Interest (ROI) Selection

The process of ROI selection for 3-dimensional platforms (Healthmyne) is as follows. First, the CT scan of interest is opened in the platform. The index slice at the level at the largest overall transverse tumor diameter is identified. The tumor is traced at this level with care to maintain the outer margins of the tracing just within the boundaries of the tumor. Once the single-slice ROI has been traced, automatic segmentation is performed. During automatic segmentation, the entire volume of tumor as seen on cross-sectional CT imaging is automatically segmented by the platform. Following automatic segmentation, the user must manually refine the tumor boundaries in order to ensure non-tumor tissues are excluded from analysis. Once correct tumor margins have been verified, the radiomics metrics are extracted. All segmentations were created by a trained medical student under the direct supervision of a fellowship trained abdominal radiologist with 11 years experience.

2.4 Data Processing and Cleaning

As discussed above in Radiomics Platform section, an initial pass was made through the data with manual exclusion of categories that were not extracting meaningful image data (coordinates etc). The data extracted from the CT scans were to the best of our knowledge and resources available, and had some missing data. To avoid data loss, we used different imputation methods to fit the data. As we were aware of no advantages to more complex methods, when imputation was needed for a method we chose a simple imputation scheme of replacing data for a component in a feature by the mean of values of other features, making use of the SimpleImputer package available in scikit-learn [20]. We made sure that no data-leakage occurred in fitting of the data by performing imputation on the training data and transforming the test data before prediction. For machine learning methods with built in imputation schemes we used those schemes. See Results (chapter 3) for more information.

2.5 Data Visualization

Before developing any model, we try to get an estimate of the data distribution and analyze if there is a clear and evident margin of classification. We visualized our data to understand the distribution over both the classes on a 2-dimensional plot using t-SNE (t-distributed Stochastic Neighbor Embedding). This is a non-linear technique used for dimensionality reduction of high dimensional data and is widely used for cancer detection applications. The data was imputed and normalized before transforming to 2-dimensional data. The t-SNE plots are shown in the Results section.

2.6 Machine Learning Analysis

The goal of our model was to evaluate for association between high nuclear grade (grade 3-4) and imaging features in our first cohort and between presence/amount of sarcomatoid features and imaging features in our second cohort. For the purpose of our study, we tested our model with gradient boosted trees (implemented in XGBoost (XGB), and Random Forest (RF) and Support Vector Machine (SVM) (implements in Scikit-learn [20]). The Scikit-learn implementations of RF and SVM do not allow for missing data while XGB has a built in imputation scheme. Therefore, we have used data imputation (see section 2.4) on training data for SVM and RF during model development. The performance of the models were evaluated on six metrics.

- **Accuracy**: This metric is the fraction of correct predictions made by the model.

$$accuracy = \frac{number\ of\ correctpredictions}{total\ number\ of\ predictions}$$

- **Precision**: Also known as positive predictive value, this metric gives the fraction of true positive predictions among total positive predictions.

$$precision = \frac{true\ positive\ predictions}{true\ positive\ predictions\ +\ false\ positive\ predictions}$$

- **Recall**: Also known as sensitivity, gives the fraction of true positive predictions among actual positive elements.

$$recall = \frac{true\ positive\ predictions}{true\ positive\ predictions\ +\ false\ negative\ predictions}$$

- **f1-score**: This metric is a measure of test's accuracy and is defined as the harmonic mean of precision and recall.

$$f1 = 2 \times \frac{precision \times recall}{precision\ +\ recall}$$

- **AUC**: Receiver operating characteristic (ROC) curve is a graph of true positive rate with false positive rate. This is a measure of a classifier's performance plotted for different classification thresholds of the classifier. The metric (AUC) is the area under the ROC curved provides an aggregate measure of the classifier's performance

- **Geometric mean**: To counter a 2-class imbalance in the dataset, another metric to determine the classification accuracy is the geometric mean score which is just the geometric mean of the true positive rate and true negative rate

We calculate the mean and standard error for these metrics using a set of values determined from 20 iterations of 5-fold cross-validated scores, where each score is determined for each fold, leading to 100 samples for each statistic being used to find mean and standard deviation.

Further, we performed permutation testing to determine the statistical significance of the model. We tested the model for f1-score averaged over 20 iterations of 5-fold cross-validation and then ran 100 random permutations of the target data pairings to input features to estimate the p-value score of the model. We observed the mean of the final model score (i.e the f1-score for every run) , and mean and standard deviation of p-values (which is a measure of fraction of how many random permutation runs performed better than the original model) over a set of 5 permutation test runs (where each run included 100 permutations, as noted just above).

2.7 Feature Ranking and Selection Strategy

The datasets had a high feature-to-sample ratio so dimensional reduction and feature selection were an important step to filter out unwanted features. We explored two approaches, including our own custom feature ranking algorithm and the algorithm used in XGB to assign a normalized importance score to each of the features.

The evaluation of features was done in two phases. In the first phase, or 'feature ranking phase', we performed 5-fold cross validation on the dataset for 20 times each time with a different train-test split. In essence, the model ran for 100 independent iterations and assigned an importance score each time. The feature list was sorted based on the cumulative importance score after 100 runs. In the second phase, or 'feature selection phase', a 5-fold CV score averaged over 20 runs was observed over the entire dataset with just the highest ranked feature from phase 1. This was repeated with the 2 highest ranked features and the average 5-fold CV score was observed. The average 5-fold CV score v/s number of features was plotted and the list of features for which the best CV score was obtained gives us the optimal set of features which we used for model optimization later. To avoid data leakage, during model assessment this feature ranking and selection was performed only for data subsets using nested cross validation, as described below.

2.8 Nested Cross Validation

As we are dealing with small datasets, a commonly known problem of data leakage often arises and can heavily impact or bias the result. To avoid this, we have evaluated the model performance using a nested cross-validation approach. We use a 5-fold loop which forms our 'outer loop'. The training data which forms the 'inner loop', in every fold, goes through the feature selection strategy described above to give the optimal feature list to be used for model optimization. We observe the scoring metrics on the test data using this optimal feature list. Each fold produces its own feature list and scoring metrics and we average the scoring metrics over the 5 folds.

2.9 Synthetic Minority Oversampling Technique (SMOTE) Analysis

For the sarcomatoid rich dataset, the samples included had a non-zero percentage of sarcomatoid features. However, a concern was that even if sarcomatoid features have an imaging signature it might be overwhelmed by the background features of the tumor if only a small percentage of sarcomatoid features are present. To address this concern and give the modeling the best chance of success, we used a filter on the percentage of sarcomatoid features present, taking only values with $>= 10\%$ (this was the median percentage in our cohort, n=25). As many of the samples had a sarcomatoid percentage less than 10%, these samples were filtered out, causing an imbalance in the samples of each class. We used SMOTE on the minority class and performed naive classification using XGB. To apply SMOTE, it was essential to use imputation (using the methods from section 2.4) before performing classification.

Chapter 3

Results

All the results obtained are enlisted below with the relevant plots and tables wherever necessary.

3.1 Patient cohorts

Two patient cohorts were evaluated. One was a group of 141 patients (46 women and 95 men, mean age 60 years) with large RCC (mean size 10±3 cm, median 9 cm) who underwent non contrast and/or portal venous phase CT used for identification of high nuclear grade (NG). This group contained mostly clear cell RCC (n=118, 84%), with fewer non clear cell (papillary n=14, chromophobe n=9). There was a slight majority of high grade tumors, (n=75 nuclear grade 3, 4) with 63 low grade (nuclear grade 1,2) and 3 tumors not graded.

The second was a group of 43 patients with RCCs with sarcomatoid features (31M, 12F, mean age 63.7 yrs) who underwent non contrast and/or portal venous phase CT with 49 size matched non-sarcomatoid RCCs from the nuclear grade cohort above to serve as controls (30M, 19F, mean age 64.4 yrs). Mean size of the sarcomatoid tumors was 9.8±3 cm, median 10 cm; mean size of controls was 8.7±2 cm, median 9 cm. Sarcomatoid tumors were predominantly clear cell (n=35, 81%). A group of 25 tumors in the sarcomatoid cohort had an estimate of the percentage of tumor with sarcomatoid features. In this subcohort, the median was 10% sarcomatoid features, mean 21%±26%, range 1-90%

3.2 Classification

We observed that XGB was the best performing classifier on each of the datasets when compared with RF and SVM. Summary of the classification results for the methods described in the section Machine Learning Analysis section above for each of the datasets with XGB is detailed in Table 3.1. Non contrast and portal venous phase CT datasets from sarcomatoid patients with size-matched controls were classified to distinguish the presence of sarcomatoid features whereas non contrast and portal venous phase CT images in patients with large RCCs were classified to distinguish the presence of high (grade 3-4) nuclear grade compared to low (grade 1-2). In the portal venous phase large RCC dataset (PV nuclear grade), the model achieved 58 % accuracy for identification of high nuclear grade, with 69 % achieved for the non contrast CT dataset (Noncon nuclear grade). In the portal venous phase sarcomatoid data set (PV Sarc), accuracy of 66 % was achieved for identifying sarcomatoid features compared to size-matched controls. For the non contrast sarcomatoid dataset (Noncon Sarc), accuracy of 60 % was obtained. We also tested using multichannel analysis on a cohort with patients from both non contrast CT and portal venous phase CT datasets and attained an accuracy of 67 %

3.3 Feature Selection

Each dataset had a different number of features and samples available for feature selection. Among 318 texture features for Noncon Sarc dataset, 463 texture features for PV Sarc dataset, 317 texture features for Noncon nuclear grade dataset and 49 features for PV nuclear grade dataset. Those that were not clinically relevant (did not describe imaging data) were manually excluded leaving 80, 85, 82 and 49 features, respectively. Our feature selection strategy then selected 3, 7, 5 and 10 radiomics features, respectively (as in Table 3.2), that were sufficient to provide a comparable accuracy to when all features were considered together. Our feature selection strategy extracted 6 radiomic features on the multichannel cohort.

Dataset	Accuracy Score	F1 Score	Precision Score	Recall Score	AUC Score	Geometric Mean Score
Noncon Sarc	0.60 ± 0.81 %	0.51 ± 1.42 %	0.57 ± 1.78 %	0.49 ± 1.71 %	0.59 ± 0.87 %	0.56 ± 1.3 %
PV Sarc	0.66 ± 0.5 %	0.62 ± 0.72 %	0.62 ± 0.95 %	0.64 ± 0.76 %	0.65 ± 0.56 %	0.64 ± 0.6 %
Noncon NG	0.69 ± 0.94 %	0.71 ± 0.9 %	0.72 ± 1.07 %	0.71 ± 1.15 %	0.69 ± 0.94 %	0.67 ± 0.99 %
PV NG	0.58 ± 0.81 %	0.60 ± 0.89 %	0.61 ± 0.85 %	0.60 ± 1.13 %	0.58 ± 0.8 %	0.57 ± 0.82 %
Noncon + PV NG	0.67 ± 0.99 %	0.69 ± 1.11 %	0.70 ± 1.03 %	0.69 ± 1.29 %	0.67 ± 1.00 %	0.66 ± 1.17 %

Table 3.1 XG Boost Model results without imputation

Dataset	Noncon Sarc	PV Sarc	Noncon NG	PV NG	Noncon + PV NG
Features	1. ENERGY HU 2. ENERGY VOXELS 3. GLCM ENTROPY	1. AUTO LARGEST PLANAR DIAMETER MM 2. NORMALIZED ABOVE MEAN DEVIATION VOXELS 3. LARGEST PLANAR ORTHO DIAMETER MM 4. SOLID VOLUME VOXELS 5. AUTO SAGITTAL SHORT AXIS MM 6. GENDER 7. SPHERICAL DISPROPORTION MM	1. ENERGY HU 2. SKEWNEDS HU 3. GLCM ROW STD 4. ENTROPY HU 5. MAX VOXELS	1. VOLUME VOXELS 2. GLCM HOMOGENITY 3. COMPACTNESS1 MM 4. TUMOR WITH RAD 5. GLCM ASM 6. GLCM COL STD 7. GLCM DISSIMILARITY 8. ENERGY HU 9. NORMALIZED ABOVE MEAN DEVIATION VOXELS 10. MEDIAN VOXELS	1. SOLID VOLUME ML 2. SOLID VOLUME VOXELS 3. GLCM DISSIMILARITY 4. COMPACTNESS1 MM 5. SKEWNESS HU 6. ENERGY HU

Table 3.2 Selected features for each dataset with XGBoost model

3.4 Nested Cross Validation

Results for our 5-fold nested CV approach (see section 2.6) are tabulated in Table 3.3. We found these average scores are comparable to classification results by 5-fold CV with XGB. PV Sarc dataset was able to achieve 67 % accuracy while the Noncon Sarc dataset accuracy was fairly low at 48 %. Noncon nuclear grade and PV nuclear grade gave similar accuracy of 56 % and 60 % respectively, while multichannel cohort gave 56% accuracy.

Dataset	Accuracy Score	F1 Score	Precision Score	Recall Score	AUC Score	Geometric Mean Score
Noncon Sarc	0.48	0.30	0.38	0.27	0.45	0.37
PV Sarc	0.67	0.64	0.64	0.65	0.67	0.66
Noncon Sarc	0.56	0.57	0.59	0.57	0.56	0.55
PV Sarc	0.60	0.62	0.63	0.63	0.60	0.59
Noncon + PV Sarc	0.56	0.55	0.61	0.51	0.57	0.55

Table 3.3 5-fold Nested CV with XGBoost model

3.5 Permutations Tests

The statistical significance of our model predictive ability was assessed by performing permutation tests averaged over 5 runs, as described in section 2.6. We used f1-score as the metric being assessed. The results are shown in Table 3.4. A low p-value score pertains to high significance of the model. Our results show that the mean p-value for each of the dataset is less than 0.10, demonstrating that the predictions are better than random with high probability.

Dataset	Mean p value	Standard deviation of p value	Average score of non permuted data
Noncon Sarc	0.08	0.04	0.65
PV Sarc	0.04	0.02	0.62
Noncon Sarc	0.04	0.03	0.64
PV Sarc	0.10	0.06	0.58
Noncon + PV Sarc	0.08	0.03	0.65

Table 3.4 Permutation test scores with XGBoost model

3.6 t-SNE Plots

The t-SNE plots for each dataset can be seen below. As we observe each of the datasets, both classes are spread across evenly and there is no clear division between them. This suggests that the input features are not strongly correlated to the aggressiveness of the RCC, consistent with the results of machine learning fitting.

For "sarc" datasets below, "sarc absent" corresponds to samples having no sarcomatoid features while "sarc present" corresponds to having sarcomatoid features. While for "NG" datasets below, "low severity NG" corresponds to samples having nuclear grade 1 and 2 while "high severity NG" corresponds to samples having grade 3 and 4.

- **noncon sarc dataset**: The dataset has a total of 64 data points with 36 belonging to having no sarcomatoid features and 28 belonging to having sarcomatoid features as shown in figure 3.1.

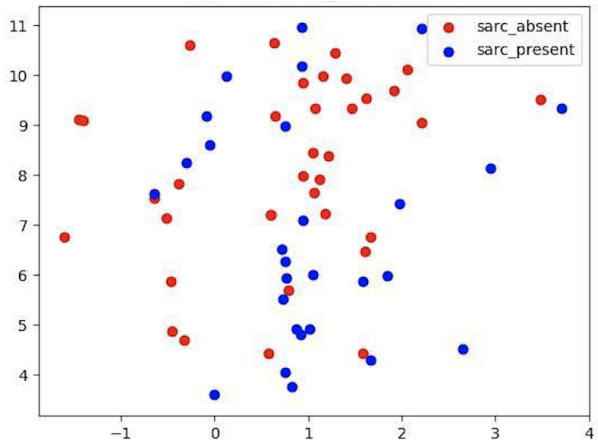

Figure 3.1 t-SNE plot for noncon sarc dataset

- **pv sarc dataset**: The dataset has a total of 89 data points with 49 belonging to having no sarcomatoid features and 40 belonging to having sarcomatoid features as shown in figure 3.2.

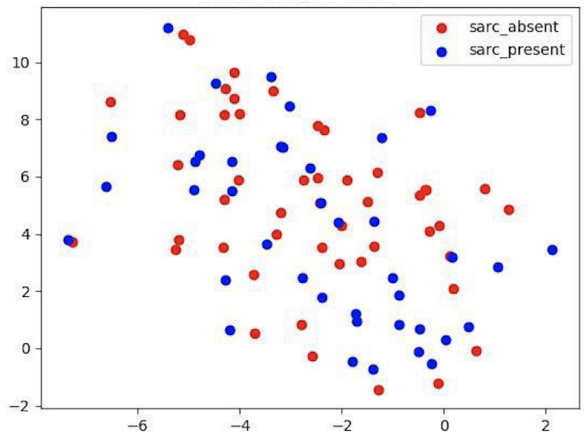

Figure 3.2 t-SNE plot for pv sarc dataset

- **pv sarc dataset with threshold**: This dataset refers to pv sarc dataset excluding data points having less than 10% sarc percentage. As shown below in figure 3.3 there were a total of 49 data points having no sarcomatoid features and 14 data points having sarcomatoid features in greater than 10% of the tumor. This minority class was oversampled using SMOTE to balance the data with majority class.

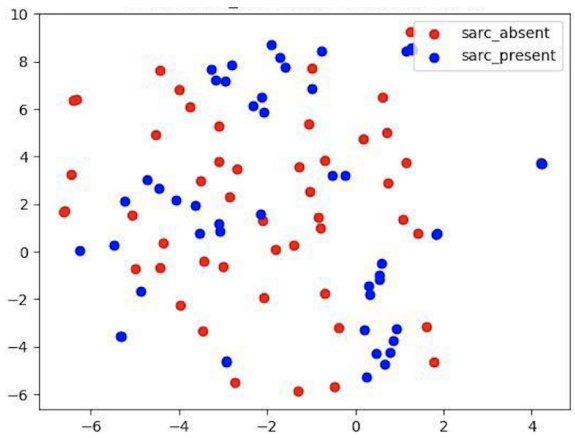

Figure 3.3 t-SNE plot for pv sarc dataset with 10 % threshold

- **noncon NG dataset**: The dataset has a total of 96 data points with 45 belonging to having nuclear grade 1 and 2 and 51 belonging to having nuclear grade 3 and 4 as shown below in figure 3.4.

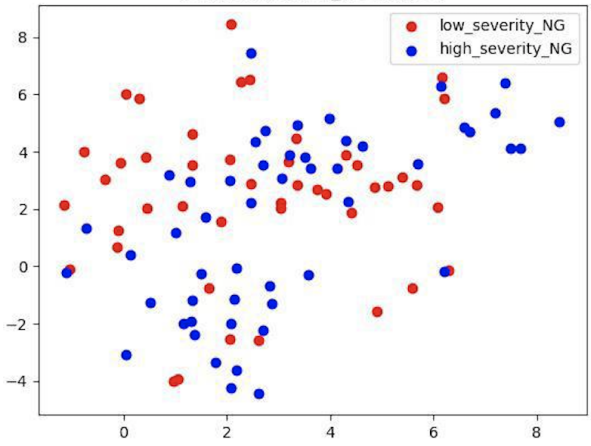

Figure 3.4 t-SNE plot for noncon NG dataset

- **pv NG dataset**: The dataset has a total of 121 data points with 56 belonging to having nuclear grade 1 and 2 and 65 belonging to having nuclear grade 3 and 4 as shown below in figure 3.5.

- **nonconpv NG dataset**: This is the dataset of patients that were present in both noncon NG and pv NG datasets. As shown in figure 3.6 the dataset has a total of 80 data points with 37 belonging to having nuclear grade 1 and 2 and 43 belonging to having nuclear grade 3 and 4.

3.7 SMOTE Analysis

For the subgroup of patients in the sarcomatoid database with percentage of sarcomatoid features included (n=25), the median was 10%. We further filtered those data samples keeping those with $>= 10\%$ sarcomatoid features to explore if tumors with more sarcomatoid features

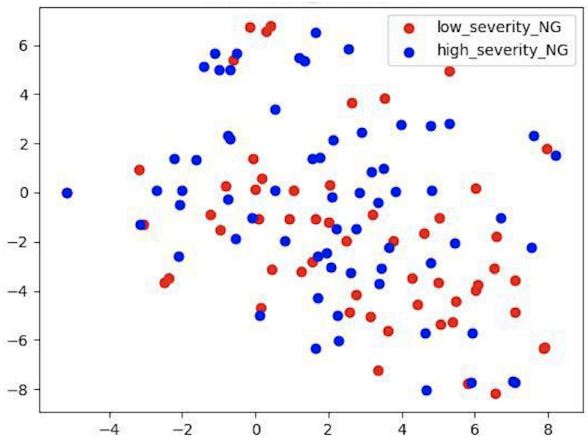

Figure 3.5 t-SNE plot for pv NG dataset

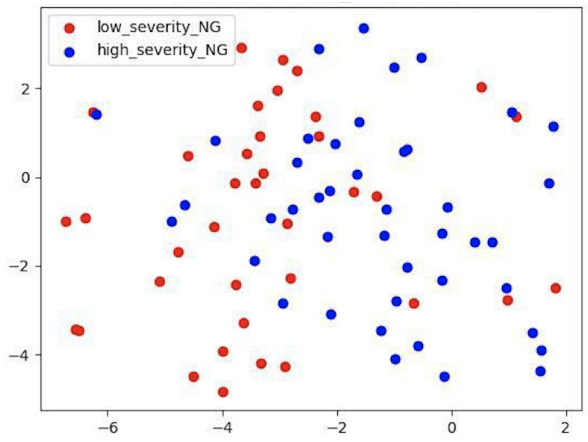

Figure 3.6 t-SNE plot for noncon + pv NG dataset

present might be better classified, as described in Sec. Synthetic Minority Oversampling Technique (SMOTE) Analysis. However, we did not observe any improvement in the classification results using XGB classifier when comparing the full subgroup of patients in the sarcomatoid database with percentage of sarcomatoid features included and those with $>= 10\%$ sarcomatoid features. The results with SMOTE are shown in Table 3.5.

Metric	Scores
Accuracy Score	0.54 ± 1.45 %
F1 Score	0.36 ± 2.72 %
Precision Score	0.46 ± 2.75 %
Recall Score	0.33 ± 2.76 %
AUC Score	0.54 ± 1.46 %
Geometric Mean Score	0.41 ± 2.6 %

Table 3.5 Classification results on PV sarc with SMOTE using 10 % threshold

Chapter 4

Discussion

Renal cell carcinoma is a heterogeneous tumor that can contain multiple different nuclear grades or genetic features in a single tumor [7]. Even if only a small portion of the tumor is grade 4 the overall nuclear grade assigned to the tumor will be 4 and that grade will drive treatment decisions and patient prognosis. Similarly, even if only a small portion of the tumor contains sarcomatoid features, if these features are identified prospectively, the identification can profoundly impact patient management and these patients are often not surgical candidates. However, sometimes these small areas may be missed at biopsy due to sampling error, and this uncertainty about the reliability of biopsy and the presence of aggressive tumor features may make prospective informed decision making about treatment challenging [2, 9, 1].

Global tumor assessment on imaging provides a non-invasive means for capturing tumor characteristics and radiomics features have shown promising associations with histopathologic features in RCC. Given the large number of radiomics features produced by many software packages, use of machine learning analysis can aid in feature extraction and robust analysis of feature and model performance. Several groups have recently looked specifically at the ability of radiomics data evaluated with machine learning to identify nuclear grade with some promising results [3, 23, 24, 8, 5, 10]. For example, Bektas et al looked at a cohort of 54 clear cell RCCs (ccRCCs), roughly half high grade tumors. They used different machine learning classifiers of 279 2D texture features extracted from portal venous phase CT. In their series, the overall accuracy, sensitivity, specificity (for detecting high grade ccRCC) and overall AUC for the best model were 85.1%, 91.3%, 80.6% and 86% respectively [3]. He et al looked at 227 ccRCCs, extracted 14 conventional imaging features manually and 556 texture features using a

software application, applied machine learning analysis, and found that the predictive models for high grade vs low grade tumors had accuracies ranging from approximately 90-94% [11].

Identification of sarcomatoid features has been challenging on CT imaging to date. Schieda et al looked at a cohort of 20 sarcomatoid RCCs matched to 25 ccRCCs and manually extracted a variety of imaging features including tumor size, subjective tumor heterogeneity, tumor margin, presence of tumoral calcification and intra and peritumoral vascularity among other features. In addition, they extracted a variety of texture features. The best performing model combined textural features and subjective features demonstrated an AUC of 0.81 in identification of sarcomatoid features [22]. Meng et al recently looked at a cohort of 29 sarcomatoid RCCs using both subjective and radiomics features and found widely variable model performance with AUCs ranging from 0.77-0.97 [26].

However, even using a similar approach in both our cohort of 141 large RCCs and 43 size matched sarcomatoid RCCs, we were unable to reproduce these results with respect to nuclear grade. We used an extensive feature selection process, applied multiple different machine learning models, tested with 5-fold nested cross validation, performed multichannel analysis of both non-contrast and pv phase post contrast data, used thresholded analysis of sarcomatoid features where quantitative data was available, and performed follow up permutation testing. There are several possible explanations for this. There is a growing body of literature that a variety of imaging parameters unrelated to biologic heterogeneity may impact selected radiomic features . In addition, there is variability in the features extracted and even the values produced for the same types of feature depending on the software platform used [15]. There have been calls for standardization to make such automatically generated features a more viable clinical tool. We also note that we used a 3D segmentation tool that incorporated the imaging features of the entire large tumor. It is possible that if only small areas of high nuclear grade or sarcomatoid features were present they may have been obscured by the dominant imaging features of the rest of the tumor. If other studies were more selective about where in the tumor the ROI was placed, or if 2D segmentation was used, this may have been less of a factor. However,

even using a threshold of 10% sarcomatoid features to select tumors with a higher percentage of sarcomatoid change, our model performance did not improve.

An additional factor that could play a role is machine learning methodology. In particular, unless great care is taken there is risk for data leakage, and the impact of even small amounts of data leakage can be significant, depending on the sample size and machine learning analysis applied. Therefore, very robust and rigorous methodology must be used. We ensured once we split the data, none of the strategies including imputation, normalization, feature selection and feature ranking were aware of any datapoint from the test data during fitting the model. Only once the model was ready, the test data was transformed as the training data before observing the prediction results. We note that by simply allowing data leakage during feature selection, we get fairly better looking results (5-7 % improvement, results shown in appendix table A.5) and this happens because test data got exposed to the model. We should therefore make sure that no data leakage happens.

The features that performed well in our model included things that make intuitive sense and are similar to those extracted in other series, including things like density, uniformity and GLCM features such as entropy. It is possible that a study using more precise radiologic pathologic correlation to look directly at the imaging features of portions of the tumors known to have aggressive features may help better delineate the imaging signature of these areas or improve model performance. This is an area of investigation that warrants further study.

There are limitations to this study. This is a relatively small dataset for this type of analysis, but it is comparable to those used in other studies, with this sarcomatoid dataset one of the largest analyzed to date. There is some heterogeneity to the CT data, but the imaging parameters used to obtain the images were within a reasonable range, and data normalization was used. Both non contrast and portal venous phase images were separately analyzed and multichannel analysis was performed where the data was available. Portal venous phase contrast was selected due to wide applicability, but other phases of contrast including corticomedullary or delayed phase images commonly used in renal imaging were not evaluated. Quantification of sarcomatoid features was only available in a subset of patients for this study.

Chapter 5

Conclusion

Despite use of a robust radiomics platform and highly effective machine learning models, performance of models for identifying aggressive tumor features in RCC (high nuclear grade, sarcomatoid features) were quite poor. Our group was unable to reproduce results seen by other groups in the literature, possibly due to variability in CT data, radiomics platforms and machine learning analyses approaches, which limits the ability to widely apply these models in clinical practice until further standardization is performed. Further study using more precise radiologic pathologic correlation may be useful in better delineating the imaging signature of these aggressive tumor features.

BIBLIOGRAPHY

[1] E. Jason Abel, Alonso Carrasco, Stephen H. Culp, Surena F. Matin, Pheroze Tamboli, Nizar M. Tannir, and Christopher G. Wood. Limitations of preoperative biopsy in patients with metastatic renal cell carcinoma: comparison to surgical pathology in 405 cases. *BJU International*, 110(11):1742–1746, 2012.

[2] Mark W. Ball, Stephania M. Bezerra, Michael A. Gorin, Morgan Cowan, Christian P. Pavlovich, Phillip M. Pierorazio, George J. Netto, and Mohamad E. Allaf. Grade heterogeneity in small renal masses: Potential implications for renal mass biopsy. *Journal of Urology*, 193(1):36–40, 2015.

[3] Ceyda Turan Bektas, Burak Kocak, Aytul Hande Yardimci, Mehmet Hamza Turkcanoglu, Ugur Yucetas, Sevim Baykal Koca, Cagri Erdim, and Ozgur Kilickesmez. Clear cell renal cell carcinoma: Machine learning-based quantitative computed tomography texture analysis for prediction of fuhrman nuclear grade. *European radiology*, 29(3):1153—1163, March 2019.

[4] Wong-Ho Chow, Susan S. Devesa, Joan L. Warren, and Joseph F. Fraumeni, Jr. Rising Incidence of Renal Cell Cancer in the United States. *JAMA*, 281(17):1628–1631, 05 1999.

[5] En-Ming Cui, Yi Lei, and Liang-ping Luo. Ct-based machine learning model to predict the fuhrman nuclear grade of clear cell renal cell carcinoma. *Abdominal Radiology*, 44, 07 2019.

[6] Giorgio Gandaglia, Praful Ravi, Firas Abdollah, Abd-El-Rahman M. Abd-El-Barr, Andreas Becker, Ioana Popa, Alberto Briganti, Pierre I Karakiewicz, Quoc-Dien Trinh, Michael A Jewett, and Maxine Sun. Contemporary incidence and mortality rates of kidney cancer in the united states. *Canadian Urological Association Journal*, 8(7-8):247–52, Aug. 2014.

[7] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012. PMID: 22397650.

[8] Shawn Haji-Momenian, Zixian Lin, Bhumi Patel, Nicole Law, Adam Michalak, Anishsanjay Nayak, James Earls, and Murray Loew. Texture analysis and machine learning algorithms accurately predict histologic grade in small (. *Abdominal Radiology*, 45, 03 2020.

[9] S. Halverson, L. Kunju, R. Bhalla, A. Gadzinski, Megan A Alderman, D. Miller, J. Montgomery, A. Weizer, Angela Wu, K. Hafez, and J. Wolf. Accuracy of determining small renal mass management with risk stratified biopsies: confirmation by final pathology. *The Journal of urology*, 189 2:441–6, 2013.

[10] Dong Han, Yong Yu, Nan Yu, Shan Dang, Hongpei Wu, Ren Jialiang, and Taiping He. Prediction models for clear cell renal cell carcinoma isup/who grade: comparison between ct radiomics and conventional contrast-enhanced ct. *The British journal of radiology*, 93(1114):20200131, October 2020.

[11] Xiaopeng He, Yi Wei, Hanmei Zhang, Tong Zhang, Fang Yuan, Zixing Huang, Fugang Han, and Bin Song. Grading of clear cell renal cell carcinomas by using machine learning based on artificial neural networks and radiomic signatures extracted from multidetector computed tomography images. *Academic Radiology*, 27(2):157 – 168, 2020.

[12] John M. Hollingsworth, David C. Miller, Stephanie Daignault, and Brent K. Hollenbeck. Rising Incidence of Small Renal Masses: A Need to Reassess Treatment Effect. *JNCI: Journal of the National Cancer Institute*, 98(18):1331–1334, 09 2006.

[13] Burak Kocak, Emine Sebnem Durmaz, Ozlem Korkmaz Kaya, and Ozgur Kilickesmez. Machine learning-based unenhanced ct texture analysis for predicting bap1 mutation status of clear cell renal cell carcinomas. *Acta Radiologica*, 61(6):856–864, 2020. PMID: 31635476.

[14] Börje Ljungberg, Steven C. Campbell, Han Yong Cho, Didier Jacqmin, Jung Eun Lee, Steffen Weikert, and Lambertus A. Kiemeney. The epidemiology of renal cell carcinoma. *European Urology*, 60(4):615 – 621, 2011.

[15] Meghan Lubner, Nicholas Stabo, E. Jason Abel, Alejandro Munoz del Rio, and Perry Pickhardt. Ct textural analysis of large primary renal cell carcinomas: Pretreatment tumor heterogeneity correlates with histologic findings and clinical outcomes. *AJR. American journal of roentgenology*, 207:W1–W10, 05 2016.

[16] Meghan G. Lubner, Andrew D. Smith, Kumar Sandrasegaran, Dushyant V. Sahani, and Perry J. Pickhardt. Ct texture analysis: Definitions, applications, biologic correlates, and challenges. *RadioGraphics*, 37(5):1483–1503, 2017. PMID: 28898189.

[17] Meghan G Lubner, Nicholas Stabo, E Jason Abel, Alejandro Munoz Del Rio, and Perry J Pickhardt. Ct textural analysis of large primary renal cell carcinomas: Pretreatment tumor heterogeneity correlates with histologic findings and clinical outcomes. *AJR. American journal of roentgenology*, 207(1):96—105, July 2016.

[18] Courtney C. Moreno, Jennifer Hemingway, Aileen C. Johnson, Danny R. Hughes, Pardeep K. Mittal, and Richard Duszak. Changing abdominal imaging utilization patterns: Perspectives from medicare beneficiaries over two decades. *Journal of the American College of Radiology*, 13(8):894 – 903, 2016.

[19] Mike M. Nguyen, Inderbir S. Gill, and Lars M. Ellison. The evolving presentation of renal carcinoma in the united states: Trends from the surveillance, epidemiology, and end results program. *Journal of Urology*, 176(6):2397–2400, 2006.

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011.

[21] Emily C. Marlow Mary Kay Theis Wesley Bolch Stephanie Y. Cheng Erin J. A. Bowles James R. Duncan Robert T. Greenlee Lawrence H. Kushi Jason D. Pole Alanna K. Rahm Natasha K. Stout Sheila Weinmann Diana L. Miglioretti Rebecca Smith-Bindman, Marilyn L. Kwan. Trends in use of medical imaging in us health care systems and in ontario, canada, 2000-2016. *JAMA*, 322(9):843–856, 2019.

[22] Nicola Schieda, Rebecca Thornhill, Maali Al-Subhi, Matthew Mcinnes, Wael Shabana, Christian Van der Pol, and Trevor Flood. Diagnosis of sarcomatoid renal cell carcinoma with ct: Evaluation by qualitative imaging features and texture analysis. *AJR. American journal of roentgenology*, 204:1013–23, 05 2015.

[23] Andrew Scrima, Meghan Lubner, E. Jason Abel, Thomas Havighurst, Daniel Shapiro, Wei Huang, and Perry Pickhardt. Texture analysis of small renal cell carcinomas at mdct for predicting relevant histologic and protein biomarkers. *Abdominal Radiology*, 44, 06 2019.

[24] Jun Shu, Didi Wen, Yibin Xi, Yuwei Xia, Zhengting Cai, Wanni Xu, Xiaoli Meng, Bao Liu, and Hong Yin. Clear cell renal cell carcinoma: Machine learning-based computed tomography radiomics analysis for the prediction of who/isup grade. *European Journal of Radiology*, 121:108738, 2019.

[25] Stuart G. Silverman, Gary M. Israel, Brian R. Herts, and Jerome P. Richie. Management of the incidental renal mass. *Radiology*, 249(1):16–31, 2008. PMID: 18796665.

[26] Li Yan, Ning Chai, Yuanzhao Bao, Yaqiong Ge, and Qi Cheng. Enhanced computed tomography–based radiomics signature combined with clinical features in evaluating nuclear grading of renal clear cell carcinoma. *Journal of Computer Assisted Tomography*, Publish Ahead of Print, 06 2020.

Appendix A:

XGB classifier performed better than RF and SVM classifiers. The tables below show the results with Random Forest classifier (RF) and support vector machine classifier (SVM):

Dataset	Accuracy Score	F1 Score	Precision Score	Recall Score	AUC Score	Geometric Mean Score
Noncon Sarc	0.64 ± 0.66 %	0.53 ± 1.00 %	0.64 ± 1.45 %	0.49 ± 1.19 %	0.63 ± 0.70 %	0.59 ± 0.84 %
PV Sarc	0.56 ± 0.85 %	0.46 ± 1.02 %	0.52 ± 1.54 %	0.43 ± 0.98 %	0.55 ± 0.85 %	0.54 ± 1.01 %
Noncon NG	0.68 ± 0.65 %	0.68 ± 0.78 %	0.71 ± 0.76 %	0.68 ± 0.99 %	0.68 ± 0.64 %	0.67 ± 0.69 %
PV NG	0.61 ± 0.71 %	0.63 ± 0.64 %	0.64 ± 0.76 %	0.64 ± 0.67 %	0.61 ± 0.73 %	0.60 ± 0.79 %
Noncon + PV NG	0.65 ± 0.62 %	0.69 ± 0.62 %	0.68 ± 0.71 %	0.65 ± 0.70 %	0.65 ± 0.65 %	0.63 ± 0.67 %

Table A.1 Random Forest model results with imputation

Dataset	Mean p value	Standard deviation of p value	Average score of non permuted data
Noncon Sarc	0.15	0.14	0.52
PV Sarc	0.30	0.17	0.44
Noncon Sarc	0.02	0.01	0.69
PV Sarc	0.05	0.04	0.65
Noncon + PV Sarc	0.07	0.10	0.67

Table A.2 Permutation test scores with Random Forest model

Dataset	Accuracy Score	F1 Score	Precision Score	Recall Score	AUC Score	Geometric Mean Score
Noncon Sarc	0.56 ± 0.75 %	0.44 ± 1.04 %	0.51 ± 1.62 %	0.42 ± 0.98 %	0.55 ± 0.74 %	0.50 ± 1.14 %
PV Sarc	0.57 ± 0.75 %	0.51 ± 0.87 %	0.52 ± 0.90 %	0.52 ± 1.06 %	0.56 ± 0.75 %	0.54 ± 0.83 %
Noncon NG	0.63 ± 0.85 %	0.65 ± 0.85 %	0.65 ± 0.84 %	0.67 ± 1.13 %	0.63 ± 0.84 %	0.61 ± 0.83 %
PV NG	0.55 ± 0.62 %	0.59 ± 0.69 %	0.58 ± 0.58 %	0.61 ± 0.89 %	0.55 ± 0.60 %	0.53 ± 0.65 %
Noncon + PV NG	0.64 ± 0.66 %	0.67 ± 0.71 %	0.67 ± 0.74 %	0.69 ± 1.08 %	0.63 ± 0.67 %	0.62 ± 0.69 %

Table A.3 Support Vector Mahine model results with imputation

Dataset	Mean p value	Standard deviation of p value	Average score of non permuted data
Noncon Sarc	0.69	0.16	0.29
PV Sarc	0.52	0.25	0.37
Noncon Sarc	0.11	0.09	0.60
PV Sarc	0.48	0.22	0.45
Noncon + PV Sarc	0.26	0.14	0.52

Table A.4 Permutation test scores with Support Vector Machine model

Dataset	Accuracy Score	F1 Score	Precision Score	Recall Score	AUC Score	Geometric Mean Score
Noncon Sarc	0.67 ± 1.03 %	0.60 ± 1.52 %	0.65 ± 1.69 %	0.59 ± 1.75 %	0.67 ± 1.09 %	0.63 ± 1.49 %
PV Sarc	0.72 ± 0.67 %	0.67 ± 0.99 %	0.71 ± 0.95 %	0.66 ± 1.12 %	0.71 ± 0.69 %	0.70 ± 0.83 %
Noncon NG	0.76 ± 0.83 %	0.77 ± 0.84 %	0.77 ± 0.81 %	0.78 ± 1.02 %	0.75 ± 0.82 %	0.75 ± 0.88 %
PV NG	0.66 ± 0.57 %	0.68 ± 0.69 %	0.67 ± 0.47 %	0.71 ± 1.09 %	0.65 ± 0.55 %	0.64 ± 0.59 %
Noncon + PV NG	0.71 ± 0.86 %	0.73 ± 0.99 %	0.73 ± 0.93 %	0.74 ± 1.26 %	0.71 ± 0.88 %	0.70 ± 1.03 %

Table A.5 XGBoost model results with data leakage