RADIOMICS ANALYSIS OF 3D COMPUTED TOMOGRAPHY IMAGES FOR PREDICTING THE ISUP GRADE OF CLEAR CELL RENAL CELL CARCINOMA TUMORS

by

Ahmet Karagöz

B.S., in Biomedical Engineering, Yeditepe University, 2018

Submitted to the Institute of Biomedical Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Biomedical Engineering

Boğaziçi University
2022

ACKNOWLEDGMENTS

I would like to thank Prof. Dr. Albert Güveniş for his continued support and guide throughout my thesis work.

Thank you to my mother for never leaving me alone.

ACADEMIC ETHICS AND INTEGRITY STATEMENT

I, Ahmet Karagöz, hereby certify that I am aware of the Academic Ethics and Integrity Policy issued by the Council of Higher Education (YÖK) and I fully acknowledge all the consequences due to its violation by plagiarism or any other way.

Name :

Signature:

Date:

ABSTRACT

RADIOMICS ANALYSIS OF 3D COMPUTED TOMOGRAPHY IMAGES FOR PREDICTING THE ISUP GRADE OF CLEAR CELL RENAL CELL CARCINOMA TUMORS

Renal cell carcinoma (RCC) constitutes %85 to %90 of all kidney malignancies. In 2020, 430,000 new cases were diagnosed and 179,000 of them lost their lives. Clear cell renal cell carcinoma (ccRCC) is the most common sub-type of RCC with approximately %80 occurrence rate. Accurate, non-invasive and preoperative determination of the International Society of Urological Pathology (ISUP) based tumor grade is important for the effective management of patients with ccRCC. Recent studies showed that CT radiomics can offer the means to predict this grade but there are some problems about data such as scarcity, unbalancing and standardization. In this study, we aimed to improve discrimination power between grades via using 3D and 2D radiomics features and ensemble machine learning methods. Radiomics features were extracted from 143 CT images obtained from the publicly available data set from The Cancer Imaging Archive. Over sampling methods and series of feature selection methods were applied to reduce the number of features. Besides the actual tumor volume, 5 additional VOIs were created to consider peritumor regions and test the robustness of the model against variations in segmentation for three ensemble machine learning algorithms. The best result was found when SMOTE was used in combination with Light Gradient Boosting Method (LightGBM) AUC of 0.89 ± 0.02. As a result, ccRCC tumor grade can be predicted from 3D CT images with a high reliability despite the inadequacy of a dataset. The algorithm is moderately robust against deviations in segmentation by observers.

Keywords: Radiomics, WHO/ISUP Grade, Peritumor, ccRCC, Machine Learning.

ÖZET

BERRAK HÜCRELİ BÖBREK HÜCRE KARSİNOMA TÜMÖRLERİNİN ISUP DERECESİNİ ÖNGÖRMEK İÇİN 3B BİLGİSAYARLI TOMOGRAFİ GÖRÜNTÜLERİNİN RADIOMICS ANALİZİ

Böbrek hücresi kanseri (BHK) tüm böbrek kanserlerinin %85 ila %90'ını oluşturur ve berrak hücreli böbrek kanseri (BHBK) en yaygın alt tipidir. 2020 yılında 430,000 yeni vaka teşhis edilmiş ve bunların 179,000'i hayatını kaybetmiştir. Tedavi sürecinin daha verimli bir şekilde yürütülebilmesi için tümör derecesinin girişimsel olmayan ve yüksek doğruluk oranına sahip yöntemlerle belirlenmesi önemlidir. Son zamanlarda yapılan çalışmalar CT görüntülerinden elde edilen radiomics özelliklerinin tümör derecelendirme işlemlerinde kullanılabileceği yönünde ama klinikte kullanılabilmesinin önünde verilerin standart olmaması ve kullanıma uygun olmaması gibi engeller var. Biz bu çalışmada 3B ve 2B radiomics özellikleri kullanarak ensemble makine öğrenmesi modelleri oluşturmayı, mevcut engellerin üstesinden gelmeyi ve daha yüksek başarımla tümör derecelendirmeyi amaçladık. Çalışmaya görüntileri The Cancer Imaging Archive'dan alınmış 143 hasta dahil edildi. Veri setindeki sınıflar arası dağılım eşitsizliğini gidermek için veri artırma yöntemlerine başvuruldu ve her hastadan çıkarılan binlerce radiomics özelliği sadece en değerli olanlar kalacak şekilde elendi. Asıl tümör alanına ek olarak 5 farkli alan daha segmente edildi ve segmentasyonda yapılacak yanlışların sonuçları ne derecede etkileyeceği araştırıldı. En yüksek başarım olan 0.89 $\pm$ 0.02 AUC, SMOTE ve Light Gradient Boosting Method (LightGBM) algoritması kombinasyonunda elde edildi. Sonuç olarak, BHBK tümör derecesini, bir veri setinin yetersizliğine rağmen yüksek güvenilirlikle tahmin edebildik ayrıca en iyi modelimizin segmentasyondaki küçük hatalara rağmen performans kaybı olmadan çalışabildiğini gözlemledik.

Keywords: Radiomics, WHO/ISUP Derecesi, Peritumor, ccRCC, Makine Öğrenmesi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D	Two Dimension
3D	Three Dimension
ADASYN	Adaptive Synthetic
ADC	Apparent diffusion coeffcient
AJCC	American Joint Committee on Cancer
AML	Angiomyolipoma
aTV	Actual Tumoral Volume
AUC	Area Under the Curve
BHK	Böbrek Hücresi Kanseri
BHBK	Berrak Hücreli Böbrek Kanseri
CA-IX	Carbonic Anhydrase IX
CCA	Correlation coefficient Analysis
ccRCC	Clear Cell Renal Cell Carcinoma
CI	Confidence Interval
CT	Computed Tomography
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
DWI	Diffusion-Weighted Imaging
eTV	Expended Tumoral Volume
FLCN	Folliculin
GLCM	Grey Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run-length Matrix
GLSZM	Gray Level Size Zone Matrix
GLV	Gray Level Variance
iTV	Inner Tumoral Volume
ISUP	International Society of Urological Pathology
KNN	K-nearest neighbor
LAD	Least Absolute Deviation

LDA	Linear Discriminant Analysis
LightGBM	Light Gradient Boosting Machine
LOG	Laplacian of Gaussian
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PN	Partial Nephrectomy
RCC	Renal Cell Carcinoma
RF	Random Forest
RN	Radical Nephrectomy
ROC	Receiver Operating Characteristic
SFS	Sequential Feature Selection
SMOTE	Synthetic Minority Oversampling Technique
SPECT	Single-Photon Emission Tomography
sTV	The Volume Surrounding The Tumor
SVM	Support Vector Machine
TNM	Tumor Lymph Node Metastasis
VHL	Von Hippel-Lindau
VOI	Volume of Interest
WHO	World Health Organization

1. BACKGROUND

1.1 Cancer Statistics

Data supplied from World Health Organization (WHO) in 2019 showed that pre-mature (0-69 years) death rates caused by cancer ranked first or second in most of the countries (112 of 183), and it also ranked third and fourth in 23 countries [1]. Approximately 19.3 million new cases were detected, and 10 million people lost their lives because of cancer in 2020 [2]. In Asia, where 59.5% of the world's population live, 50% of new cases and 58.3% of deaths happened. Even though, in Europe, where 9.7% of the world's population live, 22.8% of new cases and 19.6% of deaths were estimated to occur. America followed Europe with 20.9% of new cases and 14.2% of the death rate. More than 60% of the new cases and more than 70% of the deaths were constituted by the ten most common types of cancer. Breast, lung, collateral, prostate, and stomach cancers were the most common types with percentages of 11.7%, 11.4%, 10%, 7.3%, and 5.6%, respectively. Lung cancers and breast cancers had the highest occurrence rate and highest death rate in men and women. The distribution of cancer rates in women and men differs, and the occurrence rate in men was 19% higher than in women [1].

It is also possible to cluster countries by income; death rates caused by cancer rank second in the countries with high-income (first is cardiovascular diseases), and they rank third in countries with middle-income and low-income (first and second are cardiovascular and infectious diseases). There are internal and external factors that can cause cancer [1]. While smoking tobacco, an unhealthy diet, and radiation exposure can be given as examples of external factors, genetic mutations and hormones can be examples of internal factors [3].

1.2 Kidney Cancer

Kidney tumor, like other tumors, consists of cells that grow and spread uncontrollably. These growth changes and spread cells are formed of a mass called a renal tumor, which can be benign or malignant. Benign and malignant tumors differ in growth speed, invasion of surrounding tissues, and invasion of other body parts. While benign tumors stay in distinct areas and grow slowly without invading other body parts or surrounding tissues, malignant tumors can grow uncontrollably and spread locally or to other organs [4]. Spread to other organs or distant areas is also called metastasis.

1.2.1 Renal Cell Carcinoma

Renal cell carcinoma (RCC), Urothelial carcinoma, Sarcoma, Wilms tumor, and Lymphoma are some types of kidney cancers. The most common type of renal cancer is renal cell carcinoma which mainly grows out of the kidney's one of the essential parts responsible for filtration, also called the cortex of the kidney [5]. Kidney neoplasms accounted for 2.2% of all new diagnoses, with approximately 430,000 cases in 2020. Two hundred seventy thousand newly diagnosed kidney cancers were males, and remains were females. 1.8% of all deaths were kidney cancer patients whose number was 179,00. The number of males who lost their lives because of kidney cancer was 115,000, and 64,000 of them were females. The cumulative risk of appearance was 0.7% among males (ages 0-74) and 0.36% among females. The cumulative risk of death was 0.28% among males and 0.12% among females. Age-standardized rates of incidence were 6.1% and 3.2% in males and females, respectively. Age-standardized death rates were 2.5% and 1.2% in males and females, respectively [1].

There are innately classifications in biology, and they are among the most critical points to understand the gradual development of knowledge from class to identification and handling of diseases. With the help of the excessive amount of data published around the world, biological diversity among the existing entities has been clarified, but adding a new type or subtype has strict necessities like reproducibility and some

distinctive attributes in terms of molecular, clinical, and histopathological. For example, RCC consists of several different subtypes according to a report published by WHO in 2016 [6]. Clear cell renal cell carcinoma (ccRCC), papillary renal cell carcinoma (RCC), and chromophobe renal cell carcinoma are the most prevalent subtypes of RCC represented 70-90%, 10-15% and 3-5% of RCCs, respectively [5]. It is not easy to classify RCCs pathologically because many parameters such as location and morphological features of the tumor, genetics, and family history of patients are accounted for during classification processes. Usage of antibodies to examine if specific antigens are there in the region of interest and shape features are primarily used to characterize RCCs, although there are enormous qualification improvements at the molecular level [7]. As mentioned in the classification report [8], there is a strong correlation between tumor type and prognosis. Tumor type is also significant for the therapy that is given by following the primary therapy to increase influence, which is also called adjuvant care, besides it highlights the significance of placing tumors correctly in a category on pathological evaluation.

1.2.1.1 Clear Cell Renal Cell Carcinoma. Clear cell renal cell carcinoma is the most common subtype of renal cell carcinoma. The majority of clear cell carcinomas (95%) occur without an inherited genetic variant, and the remaining are strongly correlated with diseases such as von Hippel-Lindau (VHL) disease [5]. There are some significant discriminative features of ccRCC, such as the structure of tumors and vessels. Clear cell carcinoma cells look morphologically apparent (thanks to the high amount of glycogen and lipid) or like granular cytoplasm under the microscope. It is a rigid form of a lesion with a yellow-like color, and it can also be an alteration in the level of bleeding, internal necrosis, and corruption of cysts. Calcification might also be encountered in the tumor.

Higher levels of tumors in terms of grade and stage, the existence of necrosis, and highly aggressive behaviors as sarcomatoid do directly affect the clinical attitude of the tumor. The occurrence rate of necrosis is higher in more giant lesions and high-grade tumors [9]. It is not possible to say that ccRCC has an accurate indicator, but

immunophenotyping is feasible, and there are some markers that cause stainings, such as carbonic anhydrase IX (CA-IX) and CD10. ccRCC can be diagnosed with the help of CA-IX as the best option among markers. However, restrictions should be noticed [5]. Identifying the transparent cell carcinoma portion makes it easier to discriminate between chromophobe and high-grade clear cell carcinoma. Tests should be done for Birt-Hogg-Dubé syndrome and the folliculin (FLCN) gene germline mutation if there is a combination of chromophobe and clear cell. VHL gene mutation and Chromosome 3 Monosomy 3p are genetic mutations that widely occur in ccRCCs. Testing for VHL is not necessary for diagnosis, but germline VHL testing is suggested for patients who are younger than 46 years [9].

1.2.2 Diagnosis of Renal Cell Carcinoma

A cancer diagnosis is a complex process, and there is no single method/test that can be applied to suspected patients for precise diagnosis. For the accurate examination, the complete history of the patient in addition to diagnostic tests should be considered. Sometimes, only one test is not enough to understand whether suspected tissue is cancerous or it just imitates an indication of cancer. Besides the investigation of the existence of cancer, diagnostic tests can be used in follow-up studies to assess the influence of treatment. Therefore, tests can be repeated when patients' conditions change. Diagnostic tests include laboratory tests, diagnostic imaging, genetic tests, and biopsy.

The occurrence of RCC is often not understood with the aid of clinical signs because its clinical symptoms can take many forms, and this diversity can make diagnosis difficult and lead to misdiagnosis. Routine screening is not done for RCC diagnosis, as has been done for breast cancer in recent years, but only for those who are genetically suspected of having the disease [10]. Blood in urine, pain and a flank mass are significant symptoms for RCC diagnosis, but only a few patients have all of them. Moreover, almost half of the patients do not have any symptoms [11]. Therefore, more than half of RCC is uncovered by chance in unsuspected patients during an unrelated examination.

Because of the asymptomaticity and incidental detection, diagnosis is usually late (in advanced levels).

1.2.2.1 Genetic Tests. Genetic tests are crucial for early diagnosis of RCC because previous studies [12], [13], [14] have shown that people with a family history of RCC are at greater risk of developing RCC. In the study conducted in Sweden [13], data from all people born after 1931 was used, and it was concluded that if one sibling has RCC, the probability that the other sibling also has is very high. It was also suggested that a non-dominant gene increase the risk of RCC. Another study conducted in Iceland on 1078 patients showed that nearly sixty percent of patients with RCC have RCC in their close relatives [12]. In general, patients with sporadic RCC are diagnosed after 50, while hereditary RCC patients are diagnosed at a younger age. The emergence of RCC at a young age is another proof of hereditary transmission [15].

1.2.2.2 Laboratory Tests. Laboratory tests are standard methods used in RCC diagnosis, and also in other diseases. Laboratory tests are used to evaluate patients' situations by looking at the symptoms [11]. They do not give precise information about RCC's appearance, but they are crucial and widely used to understand whether the source of the symptoms is RCC or another malady. There are some common types of laboratory tests applied in hospitals or other health institutions, such as blood tests and urinalysis. For example, when a patient comes to the hospital with a hematuria possibility, his/her sample should be examined under the microscope several times at a few week intervals to understand the actual characteristics of symptoms. If needed, some additional tests can be applied to assess renal disease [16].

1.2.2.3 Medical Imaging. Medical imaging plays a vital role in the detection, localization, and characterization of the RCC. With the developments and explorations in the medical imaging area in the last decades, image qualities and capabilities of scanners have improved significantly. For instance, diffusion-weighted imaging (DWI)

and perfusion-weighted imaging were started to use renal mass examinations. On the other hand, despite the common usage of Positron emission tomography (PET) and Single-photon emission tomography (SPECT) in cancer detection, they are not preferred in routine clinical use [17].

Magnetic resonance imaging (MRI) was firstly used in clinics in 1980s, and its use has become more and more widespread [18]. MRI can be used as a before surgery tumor characterization tool by virtue of its ability to provide high-quality information in terms of function and structure of the tissue [19]. RCC subtypes can also be distinguished by examining the natural signal intensities of T1 and T2. DWI makes use of the motion of the water, and it is evaluated numerically with the help of apparent diffusion coefficient (ADC) maps. Cancer cells look darker in ADC images; in other words, ADC values are lower in cancerous areas because of the lower diffusion rate. In some previous studies [20], it was shown that ADC values show a great deal of variety in benign and malignant tumors. Dynamic contrast-enhanced MRI (DCE-MRI) is also a potent modality with its high resolution, so it can be used to discriminate between benign and malignant. It was also indicated that malignant tumor subtypes show different enhancements at various phases [21].

Computed tomography (CT) is one of the diagnostic imaging modalities which uses x-ray beams to create detailed images of the targeted area. Patients are placed in the CT machine before the start of scanning. X-rays are created and pass through the body (also called transmission imaging), and transmitted beams are detected by detectors during the scan. Each tissue absorbs a different number of x-rays; that is why the number of beams falling on the detector is different. Consequently, cross-sectional images are created by evaluating these absorption rates.

CT is commonly used for RCC examinations such as diagnosis, localization, and follow-up. Calcification and fat can be differentiated with the aid of CT. Thus, benign angiomyolipoma (AML), which usually consists of fat, can be detected. A calcified lesion in the kidney can be benign or malignant, and it can be differentiated via its morphology. RCC can be differentiated from renal cysts even if their densities are

high [22]. It is challenging to distinguish cancerous tissues from healthy or problematic tissue in non-contrast CT images because of the lack of enhancement. However, non-contrast images are also important in some cases because enhancements of papillary carcinoma and hemorrhagic renal cyst are confusingly similar on contrast-enhanced images [23].

In contrast-enhanced CT, the enhancement occurs in different parts of the organ or body depending on time because as time passes, the injected agent moves through the body and changes its place. Enhancement is related to the location of the injected agent and its half-life, so as time passes, the enhancement will decrease even if the agent does not move through the body. For this reason, after the contrast agent is injected into the body, it is necessary to scan within a specific time interval. The time interval at which the scanning should be completed may vary depending on the nature of the agent.

In the arterial phase, the enhancement is primarily seen in the cortical region because the contrast agent is still in the cortical capillary of the kidney while the medullary become lower enhanced. This phase is also crucial for nephron-sparing surgeries because preoperative planning becomes more manageable with the increased indistinguishability of arterial structures. Moreover, owing to a large number of the blood vessel in papillary carcinoma, enhancement is lower than clear cell RCC. That is why clear cell carcinomas can be distinguished from papillary carcinomas in the arterial phase [24]. In the nephrographic phase, the contrast distribution is homogeneous, so the enhancement is homogeneous; as a result, the contrast between the medullary and cortex is very close to each other. A previous study showed that lesion detection and characterization are easier for readers in the nephrographic phase [25]. PET/CT can also be used to classify some subtypes of RCC. However, there is a known fact that FDG uptake is very low in clear cell carcinomas, and this situation affects overall classification performance because of high incidents of clear cells in RCC.

1.2.2.4 Biopsy. Biopsy, one of the most commonly used diagnostic methods in cancer diagnosis, is invasive or minimally invasive. It is the process of taking a piece of a lesion or suspicious tissue and processing this sample pathologically. It is usually performed after laboratory tests and diagnostic imaging tests because, as a result of these tests, the suspicion of cancer increases, and the location of the lesion is determined precisely. In some biopsies, imaging techniques such as ultrasound can also be used during the procedure [26]. Percutaneous biopsy, especially fine needle aspiration, is widely used for renal tumor diagnosis and histological analysis. According to the current gold standards, the biopsy is inevitable in studies such as staging and grading because the tissue must be examined pathologically. Besides this widespread use of biopsy, it also has some problems or limitations such as accuracy, safety, pain, hemorrhage, infection, and allergic reaction. The biggest reason for some accuracy problems in RCC subtyping, staging, or grading of tumor biopsy is the heterogeneity problem in tumors. In addition, it is possible to encounter rare situations such as incorrect sampling [27].

1.2.3 Grade and Stage of Renal Cell Carcinoma

Stage and grade are essential in terms of providing comprehensive information about the size, spread, aggressiveness of the tumor. The Tumor Lymph Node Metastasis (TNM) staging system defined by the American Joint Committee on Cancer (AJCC) is one of the most widely used and has been updated over the years as findings on tumors increase. It was shown that TNM staging is vital for the prognosis and treatment plan of RCC patients. In addition, tumor grade is another critical factor in the prognosis of RCC patients and should be considered for treatment planning. Tumor grading is more complicated than staging in RCC.

Over the years, many systems have been used for RCC nuclear grading. One of the earliest was proposed in a study [28], and in this system, the grade of the tumor is determined according to the highest nuclear grade in the tissue. The grading system that emerged in the following years and has been widely used until today is the Fuhrman

grading system [29]. Although the shortcomings and inadequacies of the system were not well understood in the first years of its existence, the limitations of this system, such as intra-observer and interobserver reproducibility, began to be better understood as the studies in this field increased and the knowledge about RCC increased.For this reason, it has been questioned how valuable it is prognostically.

The latest and most comprehensive grading system is the World Health Organization (WHO) / International Society of Urological Pathology (ISUP) grading system. In this system, grading is done by considering the nuclear and nucleolar characteristics of the tumor. The WHO/ISUP grading system is as follows if the nucleoli are not visible or not evident in the 400 times magnified image, it is grade 1; if the nucleoli are visible at 400 times magnification and not at 100 times magnification, it is grade 2; if the nucleoli are visible in the 100 times magnified image, it is grade 3; and if there are extreme nuclear pleomorphism or sarcomatoid or rhabdoid differentiation or giant tumor cells, it is grade 4 [30]. WHO / ISUP tumor grade is essential in providing prognostic information and determining the possible behavior of ccRCC [5].

1.3 Radiomics

As mentioned in the previous sections, medical imaging techniques have developed significantly in recent years, even new techniques have emerged, and high quality and resolution images have begun to be obtained. However, despite all these improvements and developments, there are still situations where imaging techniques are insufficient or unreliable enough to be applied in the clinic. For example, tumor grading cannot be done without the support of invasive methods such as biopsy because some obstacles such as heterogeneity in the tumor prevent it to be done in routine clinics with only medical images. Radiomics, an area that has increased in importance and number of studies on it in recent years, can be seen as a way to overcome these obstacles. In short, radiomics can be explained as high-dimensional quantitative feature extraction from images [31]. In addition to the qualitative information, the quantitative examination of the selected or segmented area dramatically increases the amount

of information obtained from the medical image. Furthermore, thanks to these quantitative data, it is possible to make a more subjective inference or comparison [32]. The features extracted from images automatically or semi-automatically can be examined with statistical methods or machine learning methods.

1.3.1 Radiomics Workflow

The workflow of a radiomics study includes several tasks that must be done separately and sequentially; these are image acquisition, segmentation of the region of interest, extraction of features, and statistical analysis.

A patient's medical image is obtained in image acquisition, which is the first step of the Radiomics study. Various factors are affecting the image in clinical medical imaging applications, such as reconstruction algorithms, resolution of machine, and patient position [33]. Therefore, images obtained in different institutes may differ significantly. In addition, even images obtained in the same institute and under the same conditions may differ due to different reconstruction parameters. When conducting a Radiomics work, such differences should be considered.

The second stage, the segmentation stage, is vital to radiomics studies and must be accurate, reproducible, and reliable because all radiomics features are extracted from the segmented area [34]. Segmentation can be done manually, semi-automatically, and automatically, but each method has its strengths and weaknesses. Automated methods are fast but have low reliability because tissue discrimination must be high for them to work. However, this is not always possible in some tumor tissues. Manual methods are generally considered ground truth [35], but these methods are very slow, and inter- and intra-observer variabilities are high among physicians. On the other hand, semi-automatic methods are faster than manual methods and are reliable as they allow doctors to correct contours.

The third stage is feature extraction, in which the features defining the seg-

mented texture are extracted. Doing this step is essential for the correct identification of the desired tissue and directly affects the performance of the study. Images need to be preprocessed before starting feature extraction. In preprocessing, images are resampled to a specific resolution, and intensity values are divided into nominal ranges by gray level discretization. Then, features are started to be extracted from the preprocessed and standardized images. Radiomics features can be grouped into shape features, first-order statistics, second-order statistics, and higher-order statistics [36]. Shape features include properties such as volume and maximum diameter that define the shape of the segmented area. First-order statistics or histogram features express the distribution of intensity values of VOI with statistics such as mean, median, entropy, and kurtosis. Second-order statistics include inter-relationships between voxels. Grey-level co-occurrence matrix (GLCM) and Gray-level run-length matrix (GLRLM) features provide important numerical data, especially to overcome problems such as tumor heterogeneity. Higher-order statistics include features extracted from filtered images such as wavelet transform and Laplacian transforms of Gaussian filter to highlight specific structures in images.

The number of radiomics features extracted from the images reaches hundreds or even thousands, but not all features are equally important and necessary [37]. In addition, when the number of features obtained is much higher than the number of patients in the study, which is also called the curse of dimensionality, the performance of the study is negatively affected because of the probability of overfitting and false positives increase. Many feature selection methods can deal with this problem; they can also be classified as filter, wrapper, and embedded. Dimensionality reduction methods such as principal component analysis (PCA) or linear discriminant analysis (LDA) can also be used to reduce the number of features. These methods are different from feature selection methods which select some of the features because the number of features is decreased by combining or transforming some features to acquire new features. Using the selected or obtained features, classification and prediction can be made with the help of machine learning and statistical learning. K-nearest neighbor (KNN), Light Gradient Boosted Machine (LightGBM), and Random Forest (RF) are some of the powerful machine learning models.

There are some difficulties or limitations before Radiomics features can be used as a biomarker, so it does not yet fully comply with the definition of a biomarker that should be objective, accurate and repeatable [33]. Most of the CT-based radiomics features are very sensitive to change in image acquisition and reconstruction parameters. Therefore, it is questioned whether radiomics features are reproducible enough to be used in clinical studies [34]. Although some radiomics features remain stable despite differences in segmentation caused by variations of observers and algorithms, many of them are significantly affected, and the reproducibility of the study becomes challenging [34]. The number of cases involved in the study is usually much less than the number of extracted features, so feature selection or reduction methods must eliminate most of the features. Many feature selection methods can be used to avoid the curse of dimensionality problems, and selected features may vary according to the methods or combination of methods used, but there are no state-of-the-art feature selection methods. Therefore, selected features in many studies are different and this leads to a lack of reproducibility.

2. ROBUST WHOLE-TUMOR 3D VOLUMETRIC CT-BASED RADIOMICS APPROACH FOR PREDICTING THE WHO/ISUP GRADE OF A ccRCC TUMOR

2.1 Introduction

Approximately 140,000 people lose their lives due to kidney cancer each year, and over 330,000 people are newly diagnosed. Renal cell carcinoma (RCC) is among the ten most common cancer types, and RCC constitutes more than 90% of renal malignancies [38], [39]. There exist some subtypes of RCC, and the most dominant one is clear cell renal cell carcinoma (ccRCC), with an approximately 75% of occurrence rate [9].

Tumour grading plays a vital role in clinical decision making, treatment plan, and prognosis because a tumour grade indicates how abnormal the tumour cell is and how fast it is expected to grow (also called the aggressiveness level of the tumour) [40], [41]. There have been many tumour grading systems defined up to date, one of them is the Fuhrman grading system which has been widely used [42]. However, the newest and the most comprehensive grading system is the World Health Organization (WHO) / International Society of Urological Pathology (ISUP). This novel system was more robust against intra-observer and inter-observer variability problems encountered in the Fuhrman grading system [30]. In clinical practice, reducing grades to a 2-tiered grading system is common as low and high grades to increase reproducibility. Invasive methods like biopsy are widely used to obtain tumour grade. Still, there are some problems such as haemorrhages, infection, tumour spreading, contamination, and lack of information about the whole tumour tissue. The reliable tumour grade obtained by imaging methods can eliminate the need for biopsy [43].

Radiomics is an emerging method that converts radiographic images to high-dimensional minable data [44]. Radiomics produces objective and numerical features

that help to decode visually indistinguishable tissue attributes. Quantitative information extracted from images can potentially be a biomarker for tumour grade assessment because gray level size zone matrix (GLSZM), gray level co-occurrence matrix (GLCM) and gray level run length matrix (GLRLM) are known to be significantly associated with tumour heterogeneity [45], [46]. Previous studies have shown that CT-based radiomics features have predictive value for tumour grade identification [47], [48], [49] and peritumoral radiomics features can potentially have predictive performance [50].

This work aims to evaluate the performance of pre-surgical CT-based radiomics features, extracted from 6 VOIs, on discriminating low and high WHO/ISUP grades of ccRCC. VOIs are actual tumoral volume (aTV), inner tumoral volume (iTV), two expanded tumoral volumes (eTV) with different sizes, and two surrounding tumoral volumes (sTV) with varying thicknesses. Synthetic Minority Oversampling Technique (SMOTE) [51], Support Vector Machine Based Synthetic Minority Oversampling Technique (SVMSMOTE) [52], and Adaptive Synthetic Sampling Method (ADASYN) [53] were used to balance feature sets; in that way effects of up sampling methods on performance were analysed. Classifications were done by using three different ensemble machine learning (ML) models which are Light Gradient Boosting Machine (LightGBM) [54], Random Forest (RF) [55] and random subspace K-nearest neighbour (random subspace-KNN) [56]. The robustness of the classifiers against deviations in segmentation was also studied.

2.2 Materials and Methods

2.2.1 Data sets

The publicly available dataset, acquired from The Cancer Imaging Archive included presurgical 3D CT images, semantic segmentation, and clinical information of 210 patients [57], [58] was used. As per the inclusion and exclusion criteria of the study, 143 of them were found appropriate. Detailed information about the data elimination process is shown in Figure 2.1. Images of the patients whose kidneys were surgically

removed (radical nephrectomy (RN) or partial nephrectomy (PN)) were obtained during routine care between 2010 and mid-2018 at the Medical Center of the University of Minnesota. There was heterogeneity in terms of acquisition protocols and scanner manufacturers because most of the computed tomography images were acquired using different machines in years at referring institution. Figure 2.2 displays the pipeline of the radiomics based tumor grading study which includes image acquisition, preprocessing and segmentation, feature extraction and data analysis.

2.2.2 Participants

A hundred and forty-three patients with clear cell RCC were appropriate for this study. Forty-nine (34.3%) patients treated by radical nephrectomy and 94 (65.%) treated by partial nephrectomy. The average age of patients was 59.1 (range,26-86; median, 61). Ninety-three (65%) patients were male, and 50 (35%) patients were female. Forty-two (29.4%) patients had open surgery, 82 (57.3) had robotic surgery, 19 (13.3) had laparoscopic surgery. Mean pathological tumor size of the maximum 3d range was 63.6 mm (median, 48 mm; range,16.4 mm - 236.4 mm). All the cc-RCC tumors were malignant. Eighteen (12.6%) patients had tumor with WHO/ISUP grade 1, 79 (55.2%) patients had grade 2, 36 (25.2%) patients had grade 3 and 10 (7%) patients had grade 4. Grades were reduced to 2-tiered system as low (grade 1 and grade 2) and high grade (grade 3 and grade 4). Detailed information about data is shown in Table 2.1.

Table 2.1
Clinical parameters of patients.

Clinical parameter	Full cohort (n=143)	Low grade (n=97)	High grade (n=46)	p value
Age (year)	59.1 ± 12	58.2 ± 12	64.5 ± 12	0.246
Gender, n (%)				0.292
Male	93 (65%)	59 (60.8%)	34 (73.9%)	
Female	50 (35%)	38 (39.2%)	12 (26.1%)	
Smoking history, n (%)				0.81
never smoked	63 (44%)	43 (44.3%)	20 (43.5%)	
previous smoker	53 (37%)	35 (36%)	18 (39.1%)	
current smoker	27 (19%)	19 (19.7%)	8 (17.4%)	
Body mass index (kg/m2)	30.71 ± 6.32	32.16 ± 6.19	28 ± 6.53	0.312
Pathologic size (cm)	4.58 ± 3.17	3.47 ± 1.81	6.73 ± 4.15	<0.001*
Surgical procedure, n (%)				<0.001*
partial nephrectomy	94 (65.7%)	61 (62.8%)	12 (26%)	
radical nephrectomy	49 (34.3%)	36 (37.2%)	34 (74%)	
Surgery type, n (%)				0.18
open	42 (29.3%)	27 (27.8%)	15 (32.6%)	
robotic	82 (57.3%)	60 (61.8%)	22 (47.8%)	
laparoscopic	19 (13.4%)	10 (10.4%)	9 (19.6%)	

2.2.3 Image pre-processing

The distribution of the slice thickness and the pixel dimensions were between 0.5 mm and 5 mm, histograms are demonstrated in Figure 2.3. Before peritumoral and intra-tumoral region segmentation, all images were resampled to 1 mm × 1 mm × 1 mm, normalized, and discretized.

Since some texture features are very sensitive to voxel size and intensities [59], CT images and segmentations were resampled to the identical resolution. Resampling processes were performed by using the SimpleITK toolkit [60] and the B-spline interpolation technique [61] was implemented to calculate new voxel intensities. Because

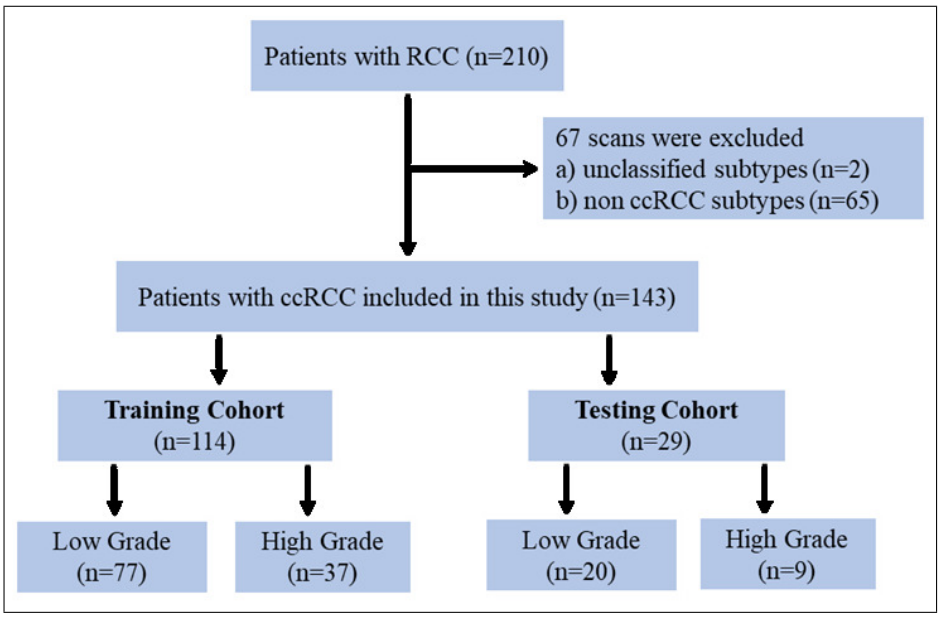

Figure 2.1 The flow diagram shows data inclusion details.

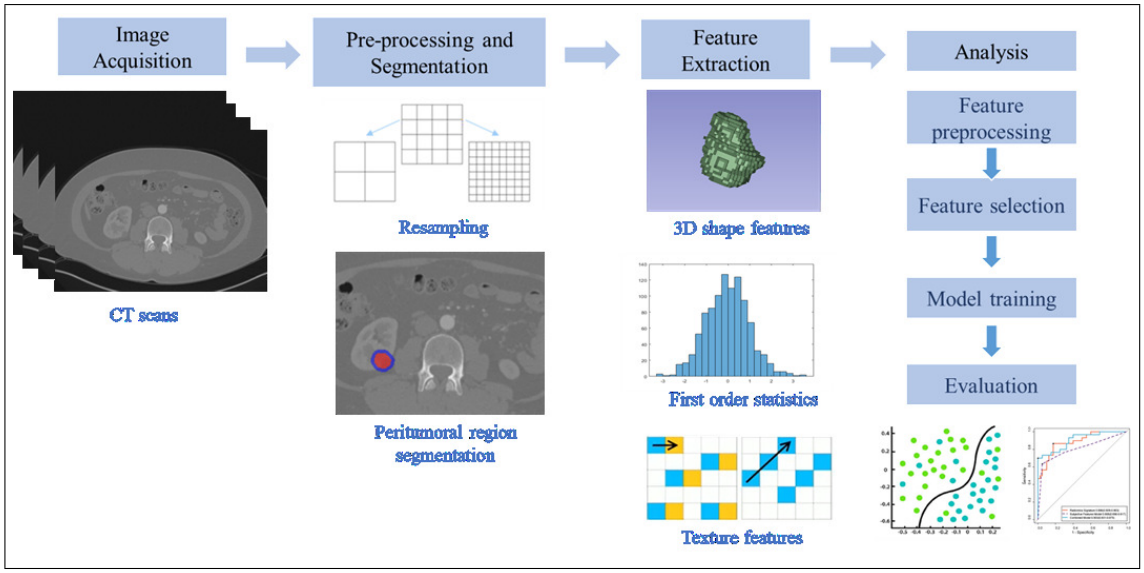

Figure 2.2 The workflow of the study.

of collecting images from various scanners, the range of intensities could be different. Therefore, the Z-score normalization [51] technique was used to eradicate inter-scanner influence.

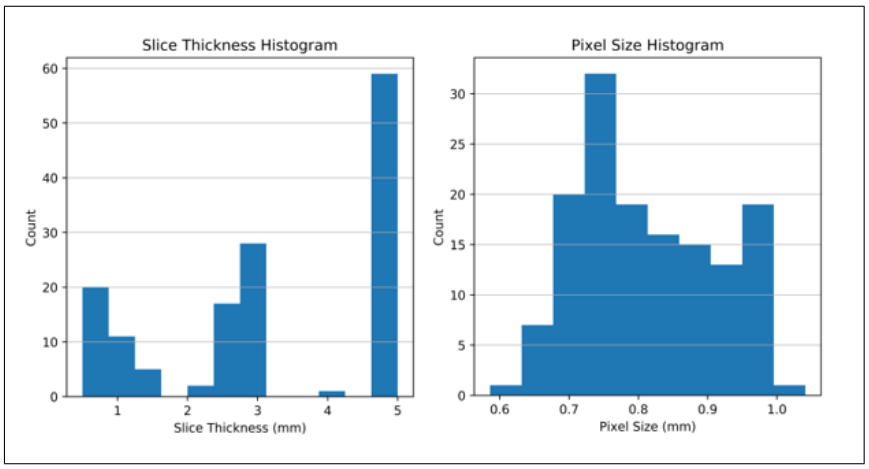

Figure 2.3 Histograms of slice thicknesses and pixel sizes before resampling.

2.2.4 Peritumoral and intra-tumoral region segmentation

In our dataset, CT images and tumour segmentations are included. We named these tumour segmentations as actual tumoral volumes (aTV). Morphological methods (also called erosion and dilation) were used to create additional regions. In addition to aTV, we created five more VOIs. Firstly, inner tumoral volume (iTV) was obtained by deleting 2mm of the aTV using the erosion method. Secondly, expanded tumoral volume with 2mm size (2mm eTV) was obtained by expanding aTV using the dilation method. Thirdly, expanded tumoral volume with 4mm size (4mm eTV) was obtained by expanding aTV using the dilation method. Morphological methods were applied in three dimensions uniformly as conducted in [50]. After getting iTV, 2mm eTV, and 4mm eTV, surrounding tumoral volumes were obtained. Surrounding tumoral volume with 2mm thickness (2mm sTV) was obtained by subtracting aTV from 2mm eTV (by taking differences of aTV and 2mm eTV). Similarly, surrounding tumoral volume with 4mm thickness (4mm sTV) was obtained by subtracting aTV from 4mm eTV (by taking differences of aTV and 4mm eTV). VOIs are represented by a sketch map in Figure 2.4.

Therefore, two evaluations have been performed:

- The effect of the peritumor region features on the classification performance
- The effect of segmentation errors in the test dataset on the total performance

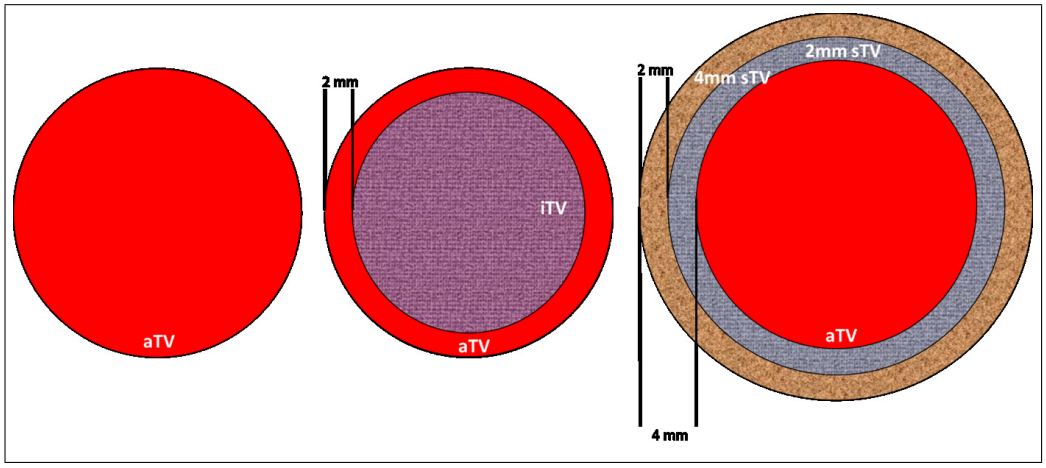

Figure 2.4 Representation of the aTVs, iTV and 2mm sTV, 4mm sTV, 2mm eTV and 4mm eTV.

2.2.5 Feature extraction

PyRadiomics [62] which is an open-source package was used to extract quantitative features from each VOI of iTV, aTV, sTV 2mm, sTV 4mm, 2mm eTV, 4mm eTV. Bin width of the gray level discretisation was adjusted to 0.01 to decrease impact of heterogeneity caused by scanning differences of scanning procedures [63]. In addition to original images, Laplacian of Gaussian (LOG) filtered, and wavelet transformed images were used to extract features. The purpose of using filtered images is to manipulate the intensity values in the image and highlight the features that are not visible or are less visible in the original images. LOG filtered images were created with values of 2, 4, and 6 to elucidate narrow, medium, and thick patterns. 3 LOG filtered images were created using values of 2, 4, and 6 to elucidate narrow, medium, and thick patterns. 8 Wavelet filtered/transformed images were created with the combination of high and low pass filters in each direction (HHL, HLH, LHH, etc.). Radiomics features are divided

into eight which are histogram or first order features (n=19), 2D and 3D shape-based features (n=26), gray-level co-occurrence matrix (GLCM) features (n=24), gray-level run length matrix (GLRLM) features (n=16), gray-level size zone matrix (GLSZM) features (n=16), neighbouring gray tone difference matrix (n=5) and gray-level dependence matrix (GLDM) features (n=14). These radiomics features were extracted from a total of 12 images (1 original, 3 LOG filtered and 8 Wavelet filtered) for each VOI, and in whole, 1133 features were obtained for each VOI.

2.2.6 Feature selection

As a pre-processing steps of feature selection, data was split into two as training (80%) and testing feature set (20%). After stratified data splitting, imbalanced training feature set contained 77 low and 37 high grade patients and testing feature set contained 20 low and 9 high grade patients. Imbalance training feature set was balanced using three different over-sampling algorithms which were SMOTE, SVMSMOTE and ADASYN. All oversampling methods were applied to the imbalanced training dataset, six times separately for all types of VOIs to compare effects of methods on predictive model training. After oversampling, imbalance training feature set were balanced to 77 instances in each class. Feature sets were standardized by using the Z-score normalization method [51] to a zero mean and unit variance in order to prevent the domination of features with wider ranges over features with smaller ranges. The training feature set was used for feature selection and model training, but the testing feature set was used only for model evaluation. To eliminate redundant features and select most discriminative features, three feature selection algorithms, namely, correlation coefficient analysis (CCA), feature importance analysis and sequential feature selection method were applied consecutively. This algorithm combination is also called the hybrid feature selection method. Firstly, CCA [64] was applied as a filter-based approach to find highly correlated features. Pearson's correlation coefficients were calculated and features with absolute correlation coefficient higher than 0.9 were eliminated. Linear support vector machine (linear-SVM) algorithm [65] with least absolute deviations loss function (also called L1 loss or LAD) was then used to calculate importance of features

and to eliminate redundant and less predictive features. Sequential feature selection (SFS) [66] with backward elimination method was finally used to choose the 15 most valuable features. Logistic regression which was the estimator of the SFS algorithm was trained using 5-fold cross validation on the training data. The hybrid feature selection process was repeated 100 times and 15 most chosen features were used for model training. After the 100 times repetition, feature selection stabilities were quantified by Jaccard index also called intersection over union [67].

2.2.7 Model training and evaluation

In order to investigate the applicability of 3D CT-based radiomics features in determining the low and high grade of ccRCC tumors, features extracted from six different VOIs were fed as input to 3 different ML models. Each imbalance radiomics feature set extracted from iTV, aTV, 2mm sTV, 4mm sTV, 2mm eTV and 4mm eTV was upsampled with SMOTE, SVMSMOTE and ADASYN and a total of 18 balanced feature sets were obtained. In addition to balanced feature sets, there were also 6 imbalance feature sets (one feature set for each VOI), therefore there were 24 training feature sets in total. 15 top ranked features were selected from 24 feature sets, and they were given as input to 3 different classifiers which were a LightGBM, Random Forest RF and random subspace-KNN. In total 72 classifiers were trained on training datasets. Testing feature sets were used to evaluate model performances by using the area under the curve (AUC) of the receiver operating characteristic (ROC) curves. Feature selection, upsampling, model training and testing processes were repeated 5 times for different training and testing partitions (5 fold nested cross validation). DeLong test [68] was used to compare ROC curves, Youden's J Statistic [69] was used to find the optimal threshold values for accuracy, precision and recall calculation.

After evaluating results of each model, performance of the model with the highest AUC value was also examined on the other feature sets extracted from iTV and 2mm eTV to interpret robustness and generalizability of the best models. Image preprocessing, feature extraction, feature selection, model training, model performance

evaluation and metric calculation processes were performed with using Python software (version 3.6).

2.2.8 Statistical Analysis

Counts (n) and percentages (%) are used to phrase categorical variables and average ± standard deviations were used to phrase continuous variables. Student's t-test and Chi-square test were used to compare continuous and categorical data. A two-tailed p-value below 0.05 was deemed statistically significant.

2.3 Results

2.3.1 Patients

In Table 2.1, distributions of patients included in this study were shown.

2.3.2 Feature selection and classification

Image resolutions were fixed to 1 mm x 1 mm x 1 mm and features were extracted from 6 different VOIs demonstrated in Figure 2.4. 1133 features were extracted from each VOI. Each feature set was split into two as imbalance train feature sets and test sets. Imbalance training feature sets were upsampled with ADASYN, SMOTE, and SVMSMOTE methods. They were used for feature selection and model training, but test sets were used only for model evaluation. Four feature sets were acquired for each VOI (1 imbalance and 3 balanced), so a total of 24 feature sets (non-up sampled feature sets included) were obtained for 6 VOIs. Feature selection steps were repeated 100 times in each feature selection process and selection stabilities were calculated, so a total of 2400 feature selection processes were performed. In each feature selection process, Pearson's correlation coefficients were calculated to find highly correlated features, and

features with a correlation bigger than 0.9 were eliminated, after elimination remaining feature numbers were between 223-258. Subsequently, a linear-SVM algorithm with L1 loss was used to find important and discriminative features. The selected number of features was between 60-82. Lastly, the most important 15 features were selected among the remaining features sequentially with the SFS method.

LightGBM, RF, and subspace KNN models were trained on training feature sets (balanced and imbalance). More than 3000 hyperparameter combinations were tried on each ML model by using the grid search method on training data to find the best hyperparameters that maximize model performance. Performance of ML models trained on 24 different training feature sets was evaluated corresponding test feature set. These processes were repeated 5 times for different training and testing partitions (5-fold nested cross validation) AUCs of each cross validated model of aTV, iTV, 2mm sTV, 4mm sTV, 2mm eTV, and 4mm eTV are shown in Table 2.2, Table 2.3, Table 2.4, Table 2.5, Table 2.6, and Table 2.7, respectively. Highest AUC values of aTV, iTV, 2mm sTV, 4mm sTV, 2mm eTV and 4mm eTV were 0.89 $\pm$ 0.02, 0.87 $\pm$ 0.03, 0.78 $\pm$ 0.09, 0.78 $\pm$ 0.07, 0.77 $\pm$ 0.06, and 0.67 $\pm$ 0.03 respectively. Highest AUC score was obtained in aTV with combination of SMOTE and LightGBM algorithm and ROC curves of each fold and mean curve are shown in Figure 2.5. ROC curves of cross validated each evaluated model are supplied in Appendix A.

The performance of the best models was also evaluated on the datasets extracted from other regions and AUCs are 0.86 $\pm$ 0.04 ($p = 0.46$) and 0.83 $\pm$ 0.05 ($p = 0.34$) for iTV and 2mm eTV, respectively.

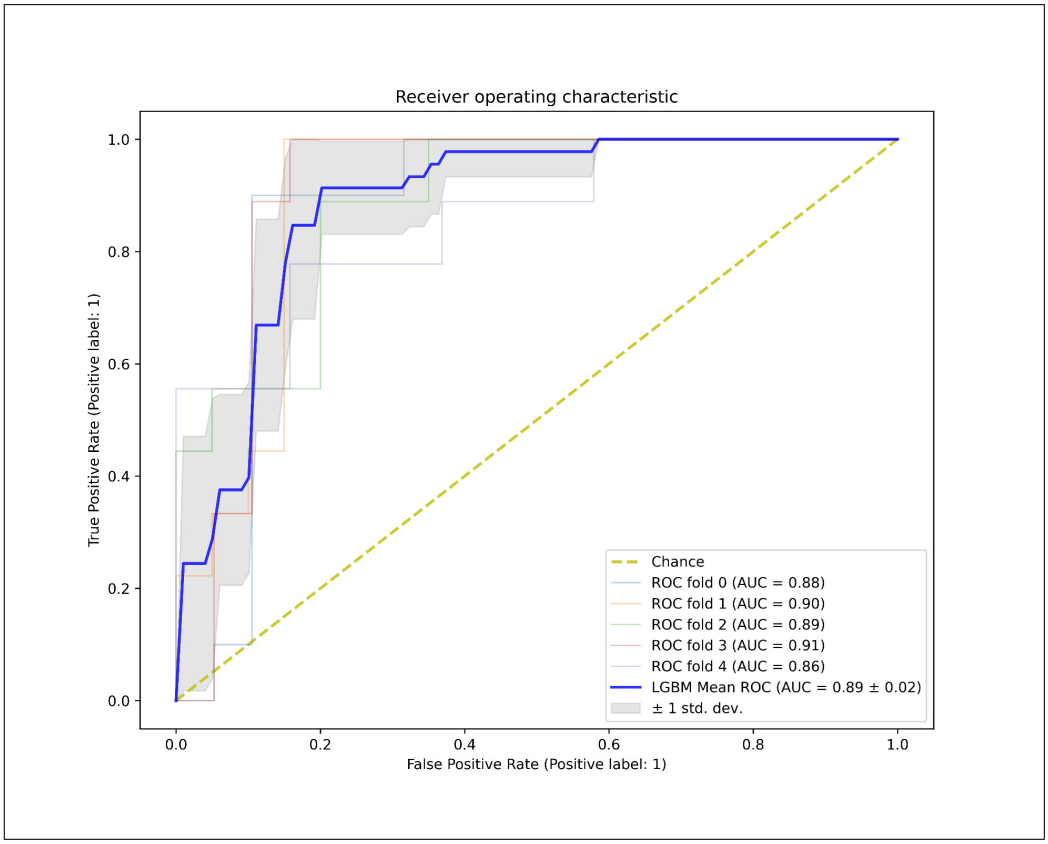

Figure 2.5 Best ROC curve of the study.

Table 2.2
Classification results of features extracted from aTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.[1]
aTV	-	0.78 ± 0.06	0.74 ± 0.09	0.74 ± 0.08	0.89
	ADASYN	0.84 ± 0.04	0.86 ± 0.04	0.82 ± 0.07	0.94
	SMOTE	0.89 ± 0.02	0.85 ± 0.06	0.80 ± 0.08	1
	SVMSMOTE	0.80 ± 0.10	0.83 ± 0.10	0.80 ± 0.10	1

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

Table 2.3
Classification results of features extracted from iTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.[1]
iTV	-	0.85 ± 0.04	0.80 ± 0.09	0.81 ± 0.09	0.82
	ADASYN	0.87 ± 0.03	0.85 ± 0.07	0.81 ± 0.09	0.76
	SMOTE	0.82 ± 0.06	0.80 ± 0.07	0.80 ± 0.08	0.93
	SVMSMOTE	0.87 ± 0.03	0.82 ± 0.10	0.82 ± 0.10	0.97

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

Table 2.4
Classification results of features extracted from 2mm sTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.
2mm sTV	-	0.62 ± 0.12	0.64 ± 0.11	0.66 ± 0.11	0.90
	ADASYN	0.71 ± 0.07	0.71 ± 0.07	0.69 ± 0.06	0.85
	SMOTE	0.73 ± 0.10	0.69 ± 0.09	0.67 ± 0.11	0.90
	SVMSMOTE	0.78 ± 0.09	0.74 ± 0.10	0.73 ± 0.09	1

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

Table 2.5
Classification results of features extracted from 4mm sTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.
4mm sTV	-	0.69 ± 0.06	0.68 ± 0.05	0.67 ± 0.06	0.76
	ADASYN	0.72 ± 0.04	0.70 ± 0.11	0.70 ± 0.11	0.99
	SMOTE	0.76 ± 0.04	0.71 ± 0.09	0.71 ± 0.08	0.80
	SVMSMOTE	0.78 ± 0.07	0.74 ± 0.08	0.72 ± 0.09	0.90

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

Table 2.6
Classification results of features extracted from 2mm eTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.
2mm eTV	-	0.74 ± 0.09	0.72 ± 0.09	0.72 ± 0.08	0.80
	ADASYN	0.77 ± 0.06	0.77 ± 0.08	0.75 ± 0.08	0.86
	SMOTE	0.69 ± 0.04	0.70 ± 0.04	0.71 ± 0.06	1
	SVMSMOTE	0.70 ± 0.08	0.71 ± 0.06	0.73 ± 0.08	0.80

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

Table 2.7
Classification results of features extracted from 4mm eTV.

		AUC			
	Up sampling	LightGBM	Subs. KNN	RF	Stab.
4mm eTV	-	0.63 ± 0.06	0.65 ± 0.08	0.66 ± 0.08	0.70
	ADASYN	0.67 ± 0.03	0.66 ± 0.05	0.66 ± 0.05	0.90
	SMOTE	0.67 ± 0.13	0.64 ± 0.11	0.65 ± 0.11	0.70
	SVMSMOTE	0.66 ± 0.07	0.66 ± 0.08	0.65 ± 0.09	0.65

[1] Stability is Jaccard's index between 0 and 1; ± 1 standard deviation.

2.4 Discussion

In this study, we investigated the predictive value of radiomics features extracted from 3D CT images for ccRCC tumour grade classification. The statistical results shown in previous section have indicated that ML algorithms can produce an encouraging performance in predicting tumour grade despite the use of small and imbalanced datasets. This can explain the high performance obtained despite the use of a small training dataset. SMOTE was also used to remedy the problem of an imbalanced dataset. Therefore, we can suggest that small and imbalanced datasets can also be used to create ML models to classify tumour grades of ccRCC. Different regions, namely, aTV, iTV, 2mm sTV, 4mm sTV, 2mm eTV and 4mm eTV were examined for WHO/ISUP based 2-tier tumour grade determination. iTV and aTV had better performance classification than sTV and eTV. Inner and outer regions of the tumorous tissue had also a predictive value, but not as much as the actual tumour volume.

Some previous studies have also shown that CT-based radiomics features had an ability to discriminate low and high grade of ccRCC. One of these studies was conducted by Feng et al. [47] on 131 patients and the feature extraction was done from 3D images, and the highest AUC value obtained in the evaluation results was 0.83. Another study was conducted by Bektas et al. [48] and 53 patients were included in the study. In this study, which was carried out by extracting features from 2D images, performance evaluation was made with the cross-validation method and the AUC value was 0.86. In the study conducted by Ding et al. [49], there were a total of 206 patients, 92 of them were in the test set, and the highest AUC they could achieve was 0.84. In addition, to the best of our knowledge, there is only one study has evaluated surrounding tissues for the ccRCC tumour grade classification. A total of 203 patients were used in the study conducted by Zhou et al. [50], and 81 of them were used as an independent test set. The features were extracted from the 3D images and the features were extracted from the peritumour area and their performance was checked. The highest AUC values are 0.848 and 0.773 for peritumoral region and tumoral mass volume, respectively. Although there was a more or less imbalance in the data sets in all these studies, an upsampling method was not used in any of them. Our first

difference from these studies is that we used 3 different up sampling methods and tried to increase model performances (the highest AUC value we obtained was obtained with one of the upsampling methods). In addition, as in several previous studies, we tried to learn about tumour characteristics by extracting features from 3D images. Another difference is that we evaluated the features extracted from a total of 6 different areas, and measured the effect of these areas on tumour grade classification. All studies are summarized in Table 2.8.

Table 2.8
Summary of previous studies and our study.

Author	**# of patients**	**ML Algorithm**	**Image**	**Grading**	**Peritumoral Assessment**	**AUC**
[47]	131	MMLR	3D / Whole slices	2-tiered / Fuhrman	NO	0.83
[48]	53	SVM*	2D / Multiple slices	2-tiered / Fuhrman	NO	0.86
[49]	206	LR	2D / Multiple slices	2-tiered / Fuhrman	NO	0.84 ± 0.08
[50]	203	LASSO	3D / Whole slices	2-tiered / Fuhrman	YES	0.85 ± 0.09
Our model	143	LightGBM*	3D / Whole slices	2-tiered / WHO/ISUP	YES	0.89 ± 0.02

In order to prevent to create biased and low performance model, up sampling methods which are SMOTE, SVMSMOTE and ADASYN algorithms were applied to training sets and model performances were evaluated on the test sets. The problem of overoptimism arises when the characteristics of the data in the training set and the test set are very similar [70]. One of the main reasons for this is the application of up sampling algorithms to the entire data set that is why only training sets were up sampled here. We applied all up sampling algorithms to improve performance, but as seen in the results, performance improvement was not always observed. All the methods used in this study make the amount of data by generating synthetic data. Synthetic data are produced similarly to the distribution of minority classes in the dataset, but in addition they create a broader and less sharp class boundary. Most of the time, this

increases the generalizability of the model and thus improves its performance. However, if synthetic data is produced very close to the class boundaries or in the overlapping areas, it can also cause performance loss [70]. The performance losses by up sampling observed in this study may be due to this reason.

As shown in Table 2.2, Table 2.3, Table 2.4, Table 2.5, Table 2.6 and Table 2.7, it is possible to conclude that up sampling methods increase the models' ability to distinguish grades. For instance, while analyzing aTV radiomics features, LightGBM's AUC score was increased from 0.78 to 0.89 with SMOTE. Combination of SMOTE and, LightGBM performed best on the aTV, but performance of combination of ADASYN and subspace KNN is also very close to best. However, it is not possible to compare performance of only up sampling methods or only ML algorithms, because combination matters. For example, while looking at the performance of LightGBM on aTV feature assessment (see Table 2.2), SMOTE is better than ADASYN, but ADASYN performs better with subspace KNN than SMOTE. Similarly, comparing performance of LightGBM and subspace KNN is not possible because subspace KNN has higher AUC score with ADASYN, but LightGBM has higher AUC value with SMOTE. Therefore, it is possible to make inferences that combinations of up sampling methods with LightGBM and subspace KNN performs better than other cases and there is no best up sampling method, but balanced training dataset give better results than original unbalanced dataset. Feature selection stability of the datasets up sampled with SVMSMOTE is higher than the other most of the time, not like SMOTE algorithm uses KNN to create synthetic data to up sample minority class, SVMSMOTE uses support vectors to find borders between classes and create new synthetic data in class region, so new dataset's data distribution in features are more discrete than dataset created with SMOTE [52] and this results with higher in feature selection stability.

When looking at the best results of each VOI (see Table 2.9), aTV performed best for ccRCC tumor grade classification. Best performed models of each VOI are listed in Table 2.9 and p values were calculated to compare whether they are significantly different from best model of the aTV. According to p values, there is a significant difference only between aTV and 4mm eTV with p=0.008 (significance level is 0.05).

aTV and iTV gave almost same results (p=0.77) with AUC 0.89 ± 0.02 and 0.87 ± 0.03, respectively, so incomplete segmentation of the tumor area, in other words segmenting not the entire tumor but a big portion of it, is sufficient to understand tissue properties and classify tumor grade with near full performance. Therefore, while performing tumor segmentation, it should be paid attention to whether the selected areas are outside the tumor area, because the properties of the tissues outside the tumor are very different from the tumor and this error in segmentation can unfavourably affect radiomics and their prediction performance. It should also be noted that because peritumour segmentation is performed automatically, peritumour areas may contain different non-tumor tissues, and therefore, features extracted from peritumour areas may not represent the tissue properties of the tumor with perfect accuracy. Furthermore, the heterogeneity of data acquisition and scanner conditions may also play a role.

However, the performance of the best model which was trained on the features extracted from aTV using the features extracted from iTV and 2mm sTV shows us that AUC values are not significantly affected, in other words, the best model is robust and generalizable enough to compensate 2mm errors in segmentation.

Table 2.9
Best performed model of each VOI

VOI	Upsampling	ML	AUC	ACC	RECALL	PREC	p value
aTV	SMOTE	LightGBM	0.89 ± 0.02	0.84	0.83	0.86	-
iTV	SVMSMOTE	LightGBM	0.87 ± 0.03	0.81	0.80	0.82	0.77
2mm sTV	SVMSMOTE	LightGBM	0.78 ± 0.09	0.72	0.71	0.76	0.16
4mm sTV	SVMSMOTE	LightGBM	0.78 ± 0.07	0.71	0.70	0.75	0.18
2mm eTV	ADASYN	LightGBM	0.77 ± 0.06	0.70	0.69	0.73	0.09
4mm eTV	ADASYN	LightGBM	0.67 ± 0.03	0.65	0.64	0.68	0.008*

The tumour grade of ccRCC is significant in determining a patient's prognosis and a treatment plan, so this study may have clinical and practical impacts as these grade estimates are obtained pre-operatively and non-invasively. Invasive methods

like biopsy can be time consuming, biased and non-reproducible because of tumor heterogeneity, but radiomics assessments can be a good method for ccRCC tumor grading rapidly and accurately before surgery.

Although our ensemble models were trained and tested on a small dataset obtained from a single center, they had ability to discriminate tumor grade. However, there are some limitations to our study. Larger and multicentric datasets are needed to assure more generalizable models.

In conclusion, this work presents a method to obtain robust and accurate WHO / ISUP grade prediction of ccRCC tumours preoperatively and noninvasively, even with small and imbalanced datasets. To achieve this radiomic features were extracted from 3D whole tumour images, proper up sampling methods were applied, hyperparameters were optimized and appropriate ensemble learning algorithms were used. With these techniques, the change in performance may not be very significant with operator induced deviations in segmentation. Further research is needed to confirm these results and evaluate the performance with manually drawn peritumour as well as different or larger databases.

3. List of publications produced from the thesis

1. Robust whole-tumor 3D volumetric CT-based radiomics approach for predicting the WHO/ISUP grade of a ccRCC tumor, A. Karagöz, A. Güveniş, *"Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization"* (Under review)

2. Predicting the Grade of Clear Cell Renal Cell Carcinoma from CT Images Using Random Subspace-KNN and Random Forest Classifiers A. Karagöz, A. Güveniş, *2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 13–14 May, 2021,.

APPENDIX A. ROC CURVES AND SELECTED FEATURES

A.1 ROC Curves and Selected Features of aTV

Performance of three ML models trained with imbalance feature sets extracted from aTV are shown in Figure A.1. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from aTV are shown in Figure A.2. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from aTV are shown in Figure A.3. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from aTV are shown in Figure A.4. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

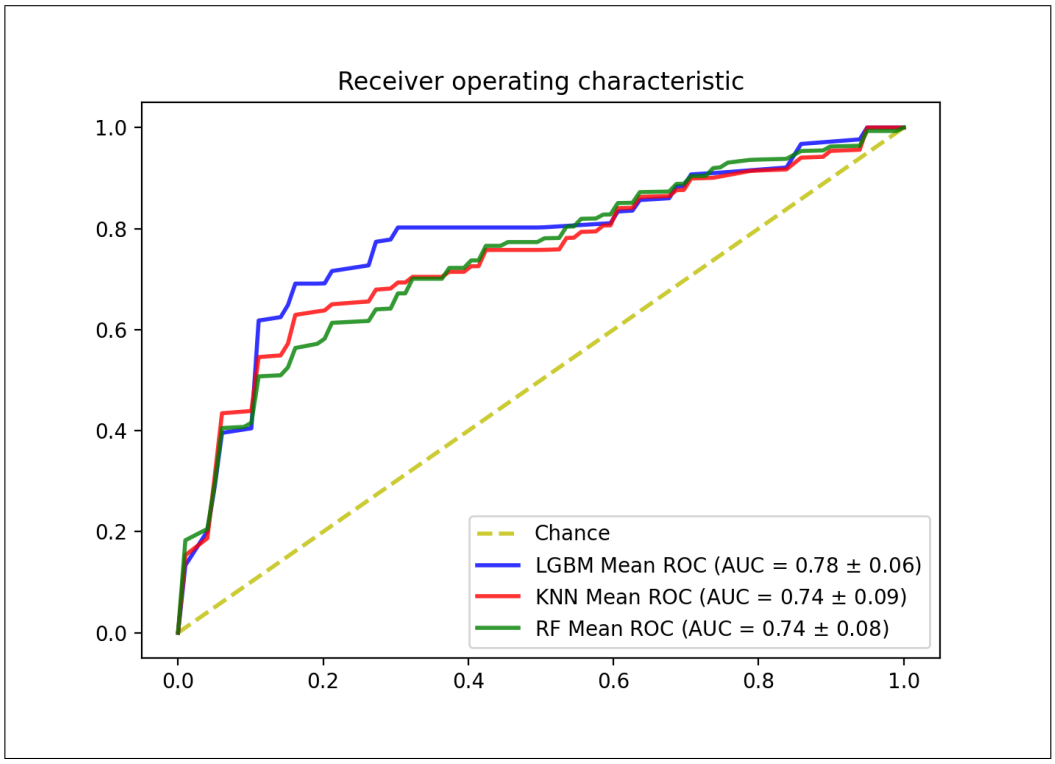

Figure A.1 Performances of three models trained with original features obtained from aTV.

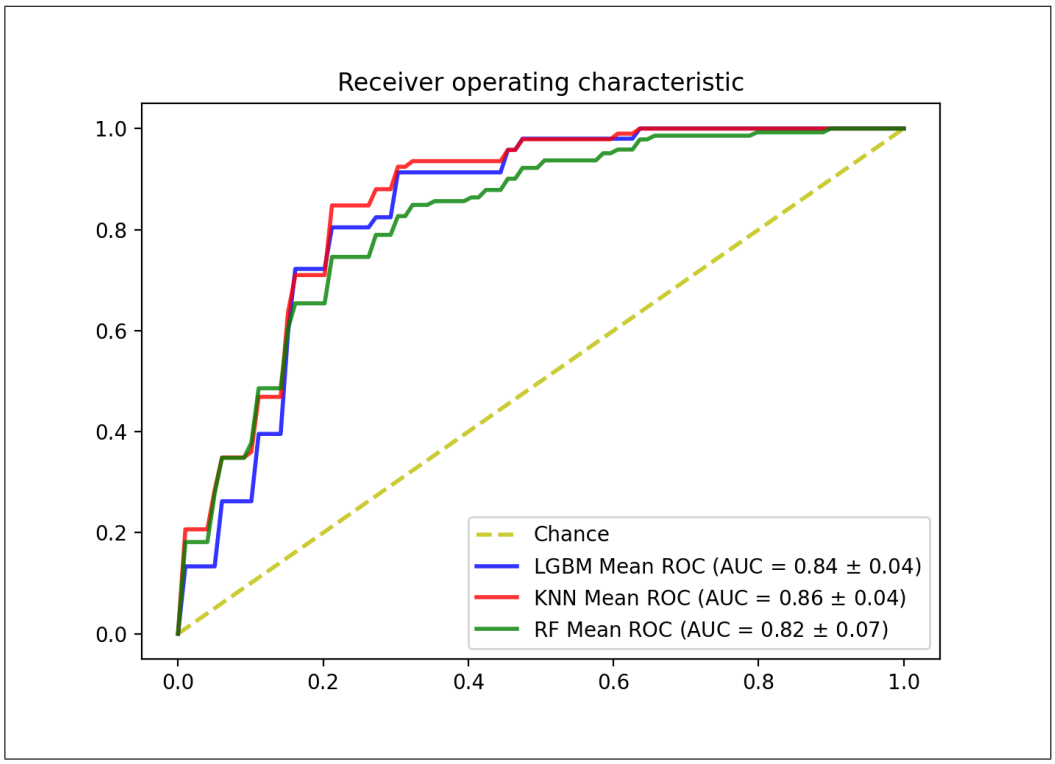

Figure A.2 Performances of three models trained with up-sampled (ADASYN) features obtained from aTV.

A.2 ROC Curves and Selected Features of iTV

Performance of three ML models trained with imbalance feature sets extracted from iTV are shown in Figure A.5. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from iTV are shown in Figure A.6. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from iTV are shown in Figure A.7. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from iTV are shown in Figure A.8. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

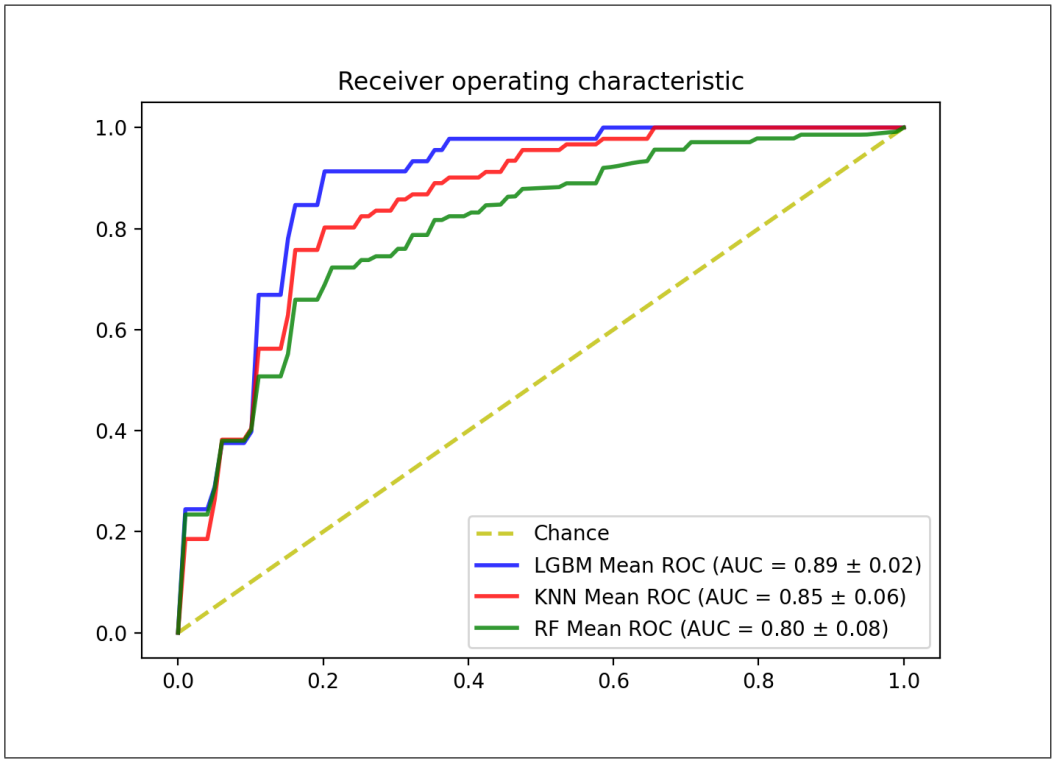

Figure A.3 Performances of three models trained with up-sampled (SMOTE) features obtained from aTV.

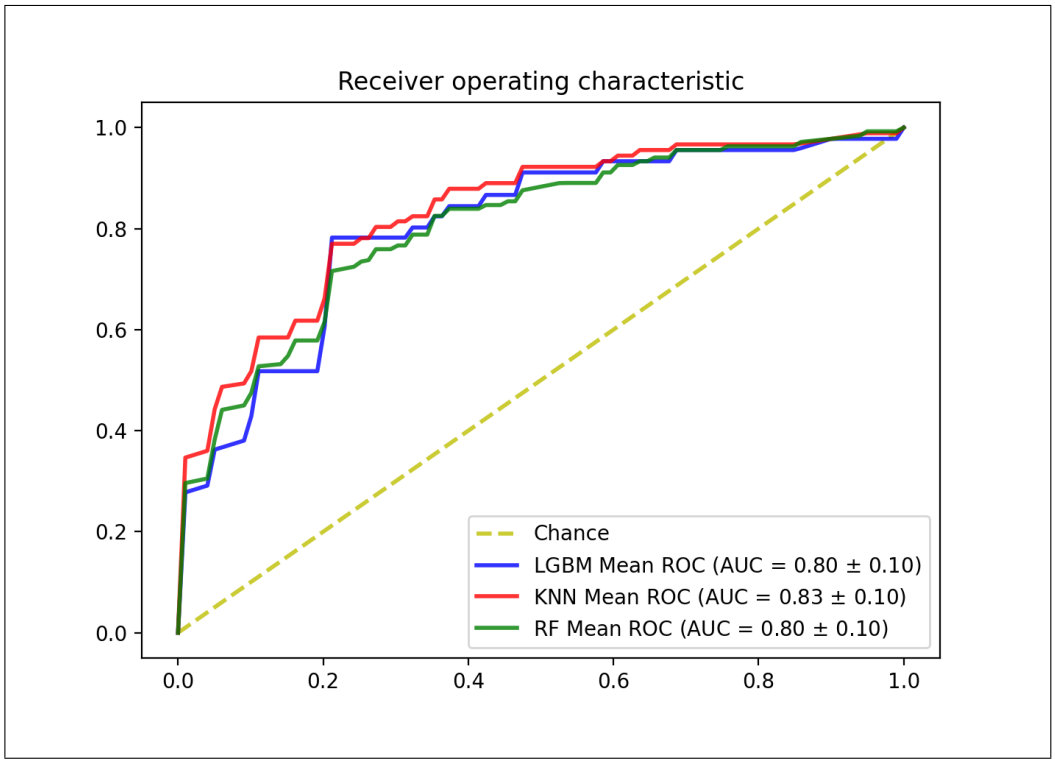

Figure A.4 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from aTV.

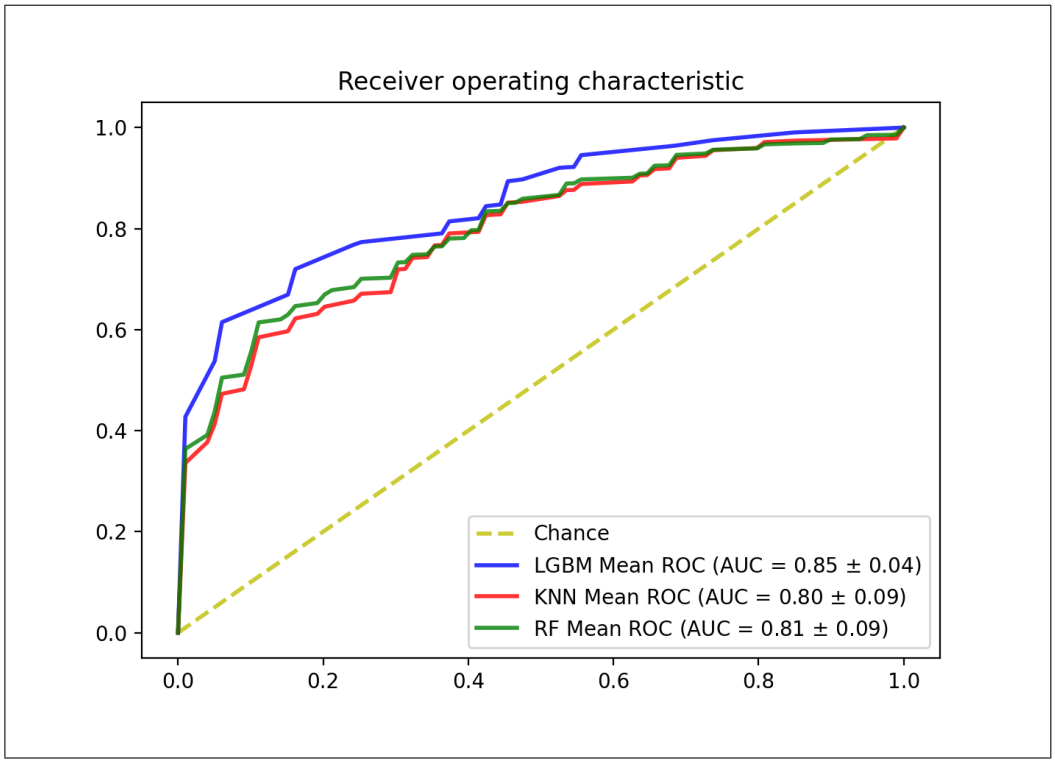

Figure A.5 Performances of three models trained with original features obtained from iTV.

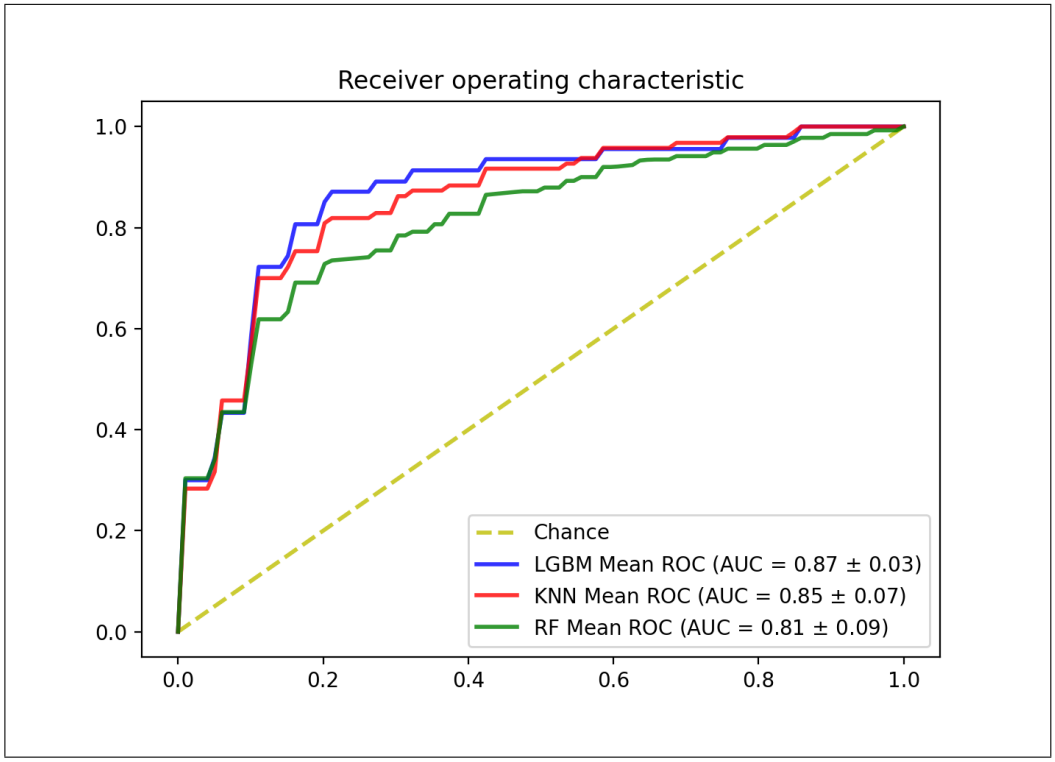

Figure A.6 Performances of three models trained with up-sampled (ADASYN) features obtained from iTV.

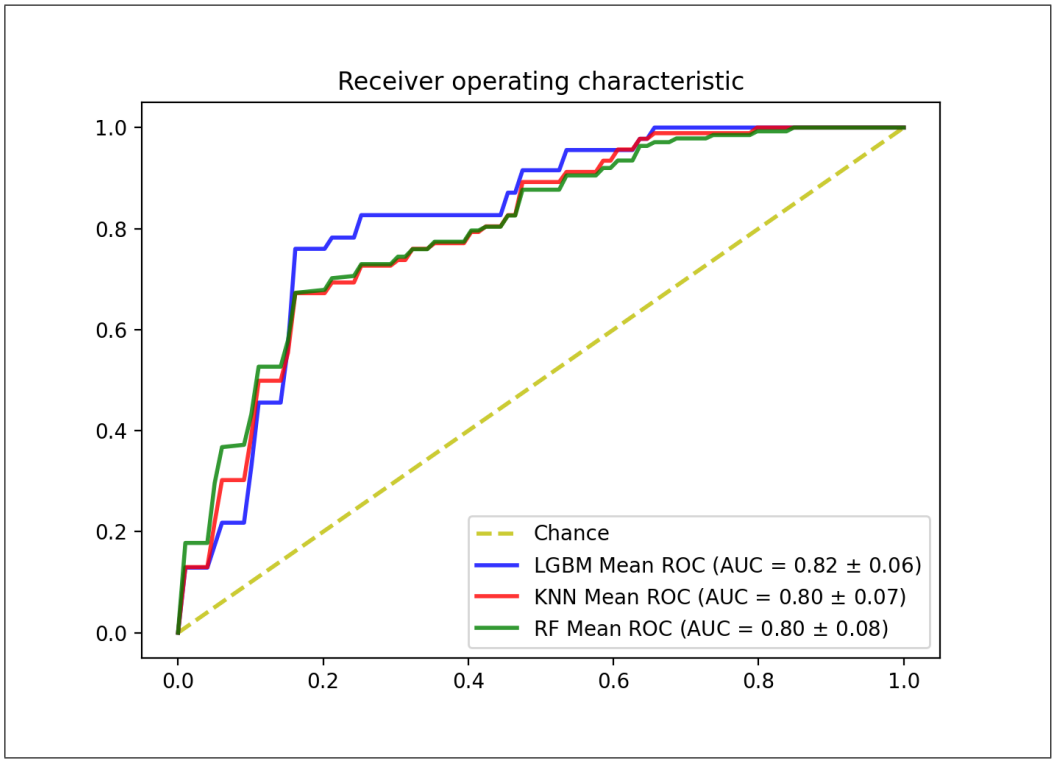

Figure A.7 Performances of three models trained with up-sampled (SMOTE) features obtained from iTV.

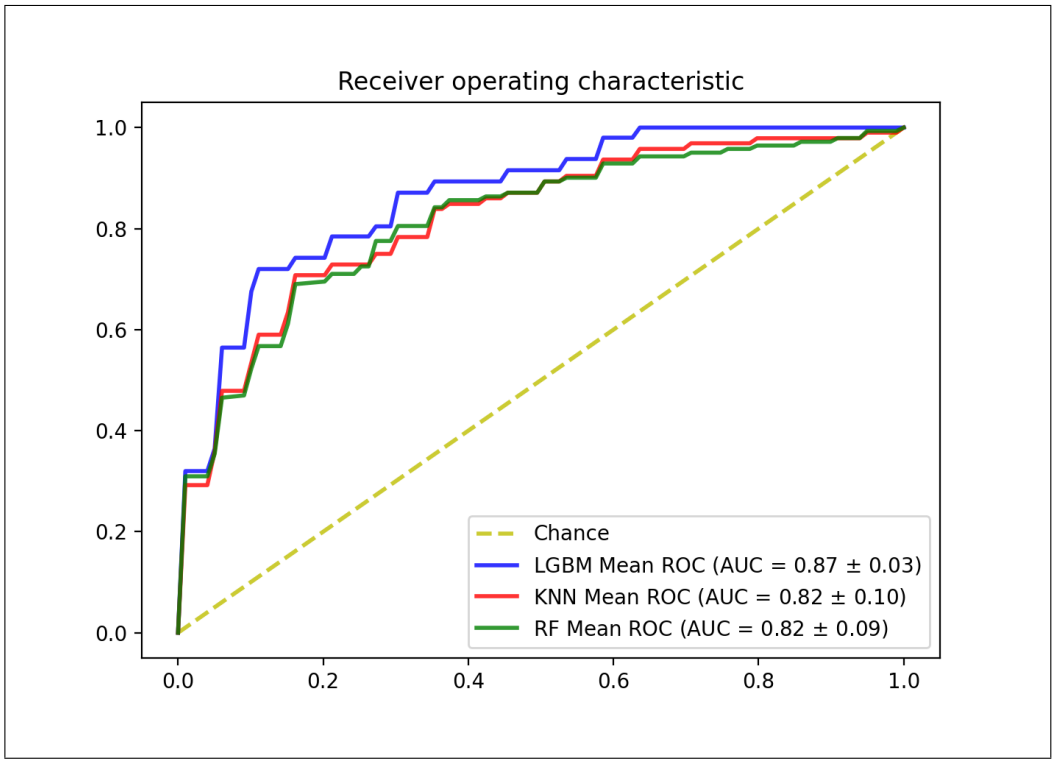

Figure A.8 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from iTV.

A.3 ROC Curves and Selected Features of 2mm sTV

Performance of three ML models trained with imbalance feature sets extracted from 2mm sTV are shown in Figure A.9. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from 2mm sTV are shown in Figure A.10. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from 2mm sTV are shown in Figure A.11. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from 2mm sTV are shown in Figure A.12. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

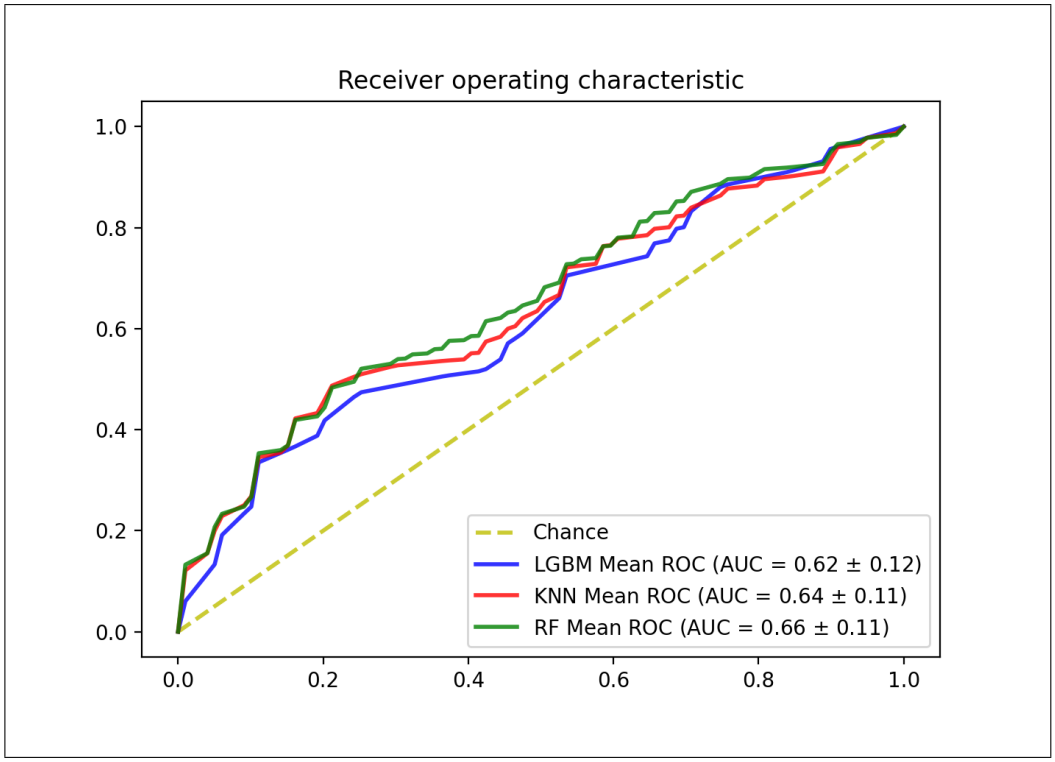

Figure A.9 Performances of three models trained with original features obtained from 2mm sTV.

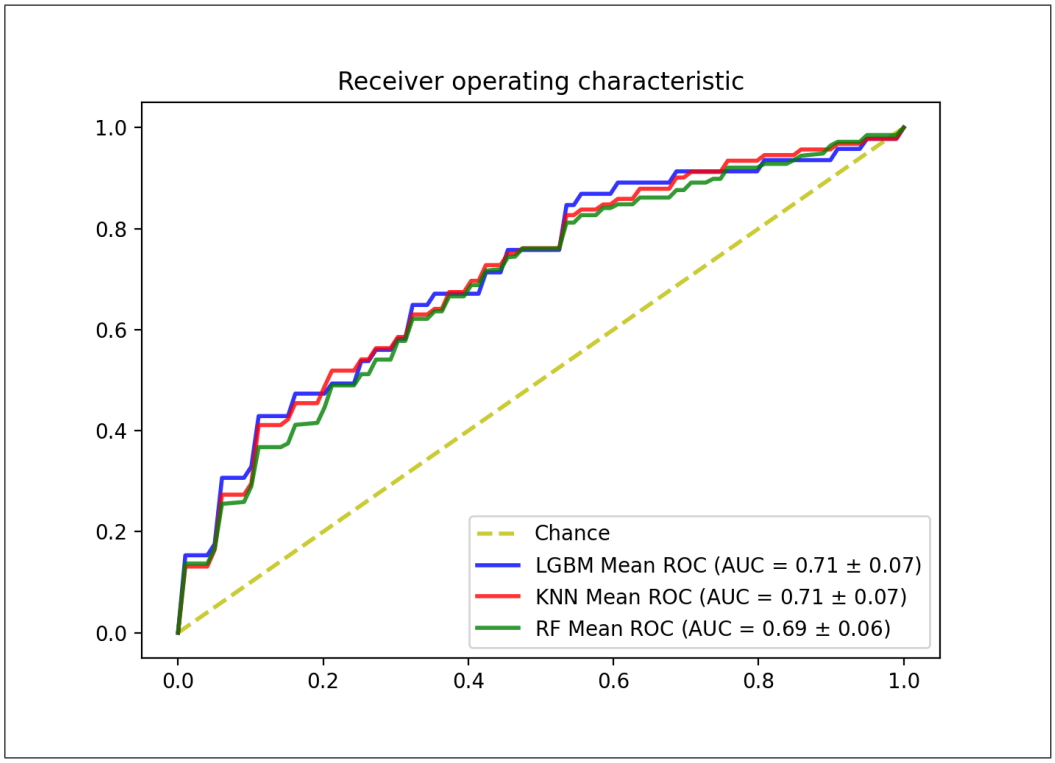

Figure A.10 Performances of three models trained with up-sampled (ADASYN) features obtained from 2mm sTV.

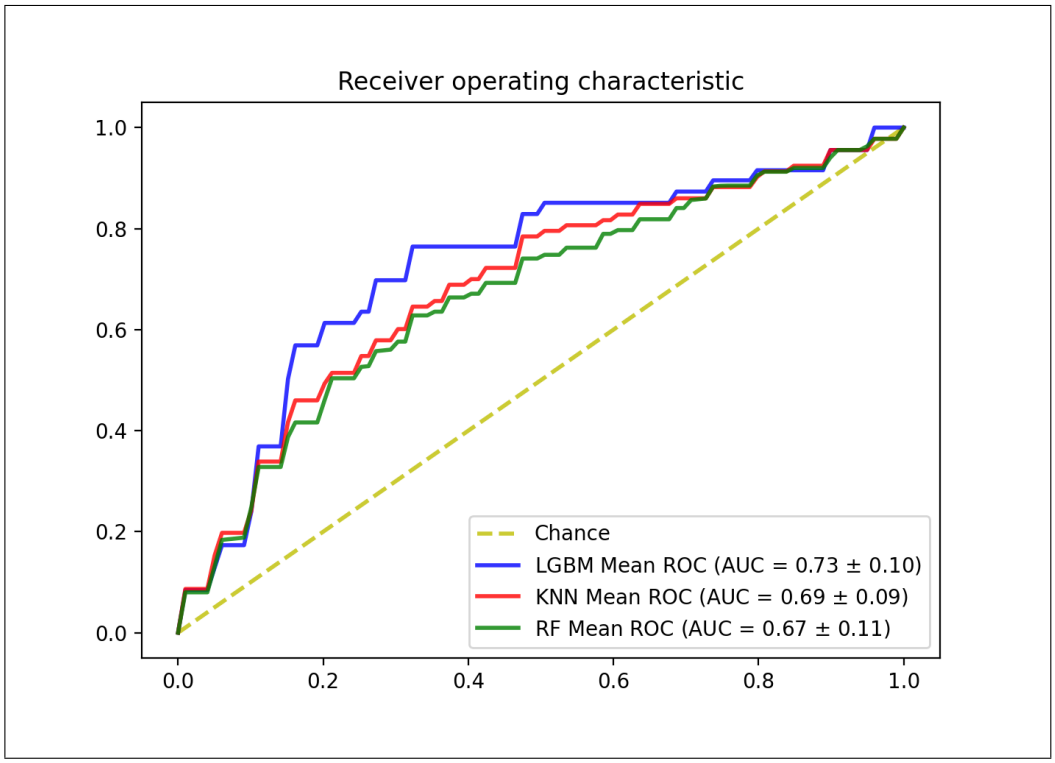

Figure A.11 Performances of three models trained with up-sampled (SMOTE) features obtained from 2mm sTV.

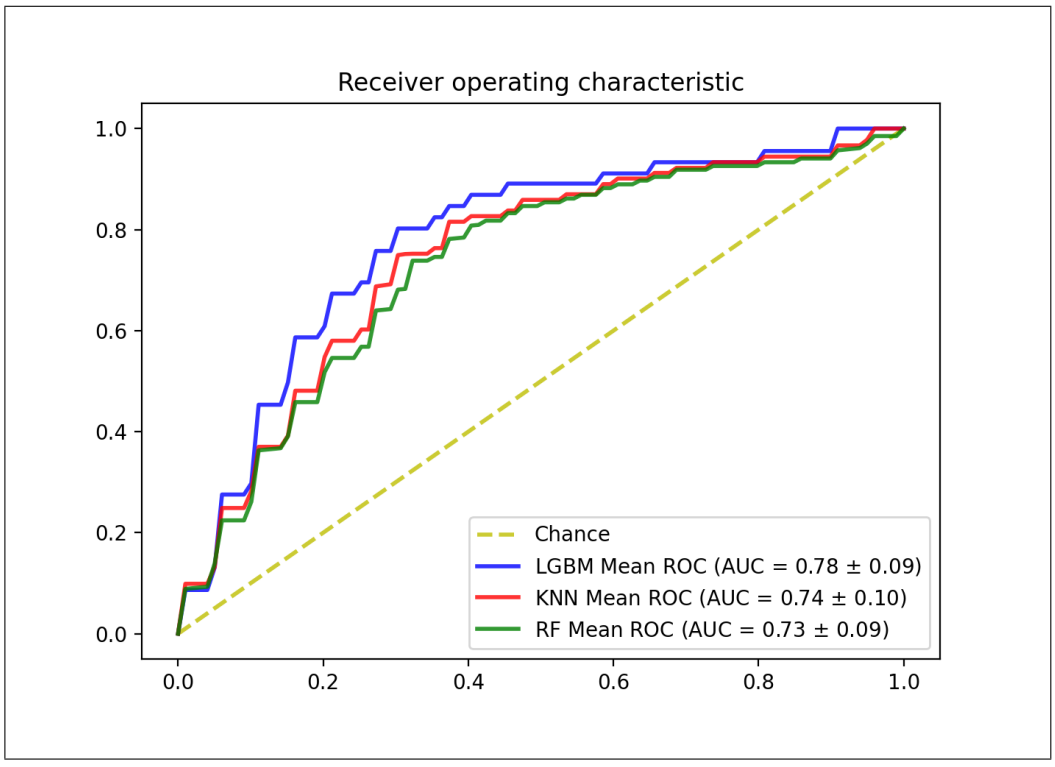

Figure A.12 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from 2mm sTV.

A.4 ROC Curves and Selected Features of 4mm sTV

Performance of three ML models trained with imbalance feature sets extracted from 4mm sTV are shown in Figure A.13. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from 4mm sTV are shown in Figure A.14. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from 4mm sTV are shown in Figure A.15. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from 4mm sTV are shown in Figure A.16. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

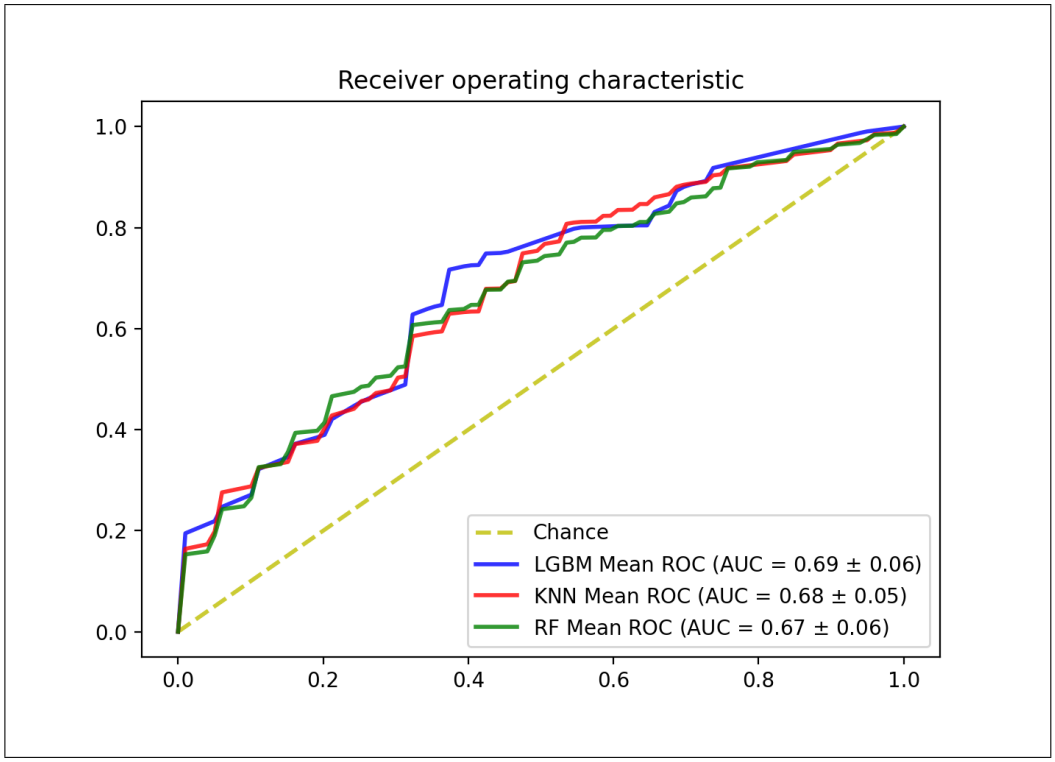

Figure A.13 Performances of three models trained with original features obtained from 4mm sTV.

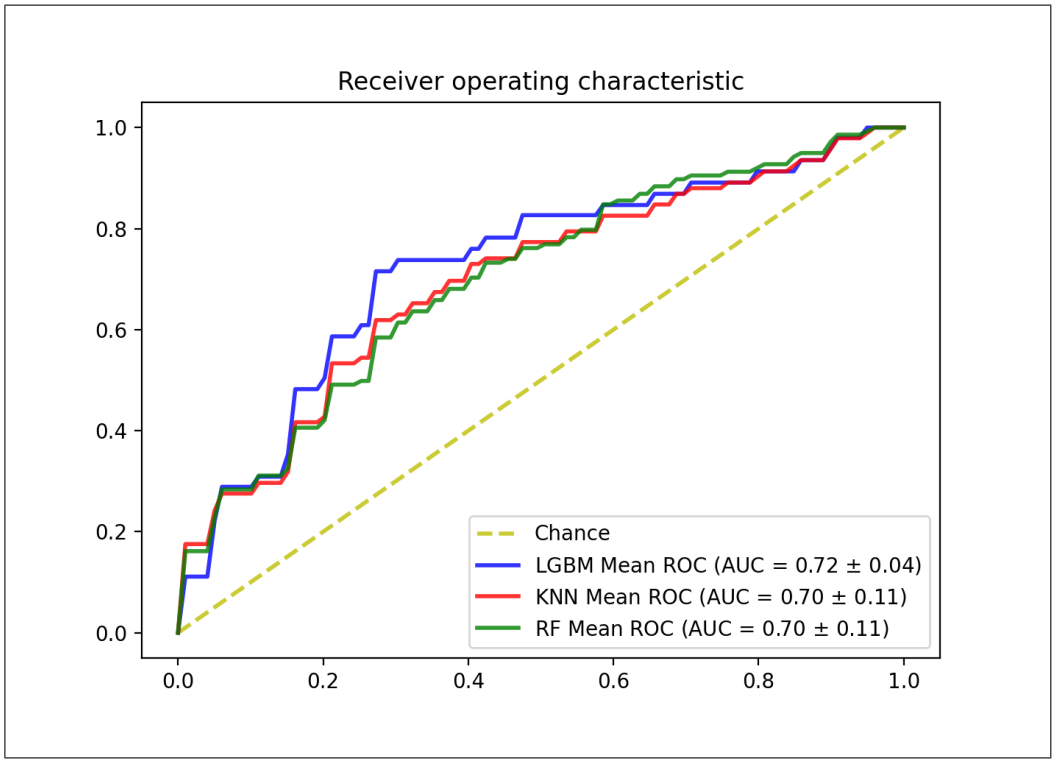

Figure A.14 Performances of three models trained with up-sampled (ADASYN) features obtained from 4mm sTV.

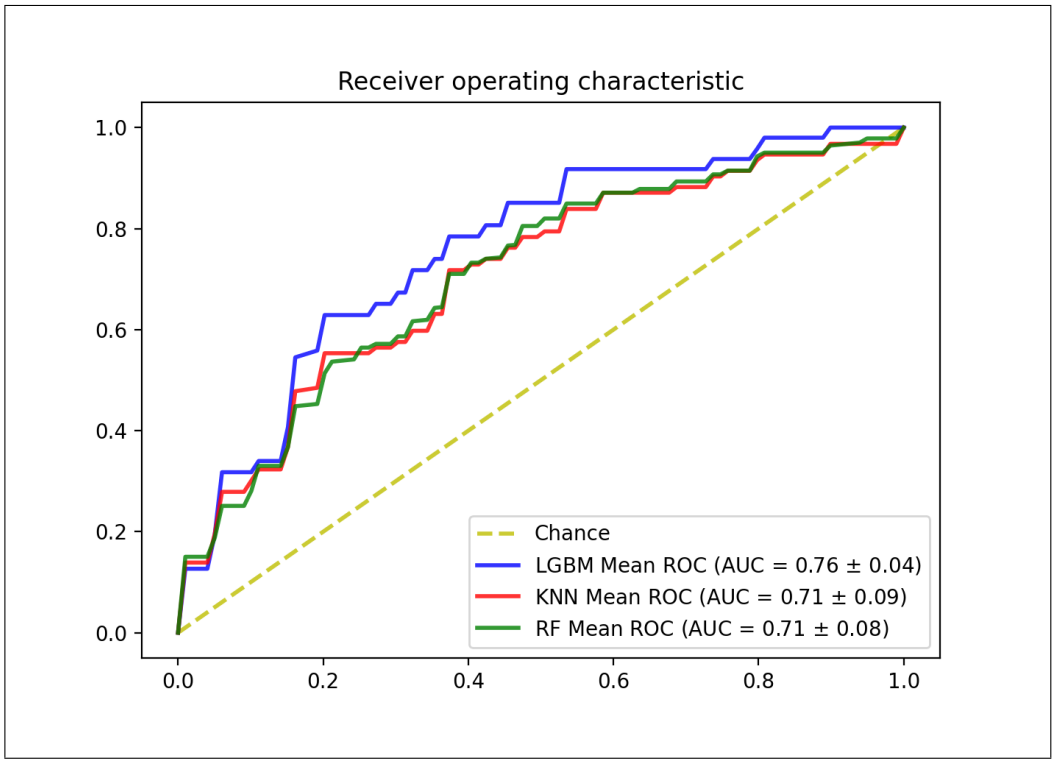

Figure A.15 Performances of three models trained with up-sampled (SMOTE) features obtained from 4mm sTV.

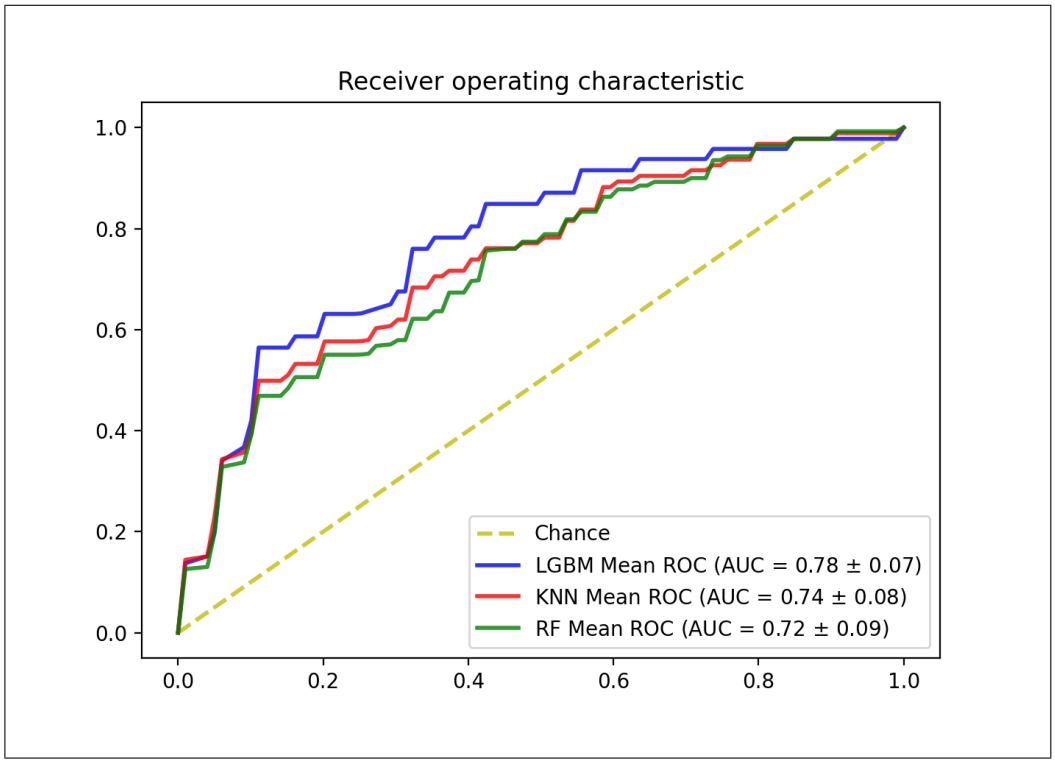

Figure A.16 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from 4mm sTV.

A.5 ROC Curves and Selected Features of 2mm eTV

Performance of three ML models trained with imbalance feature sets extracted from 2mm eTV are shown in Figure A.17. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from 2mm eTV are shown in Figure A.18. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from 2mm eTV are shown in Figure A.19. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from 2mm eTV are shown in Figure A.20. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

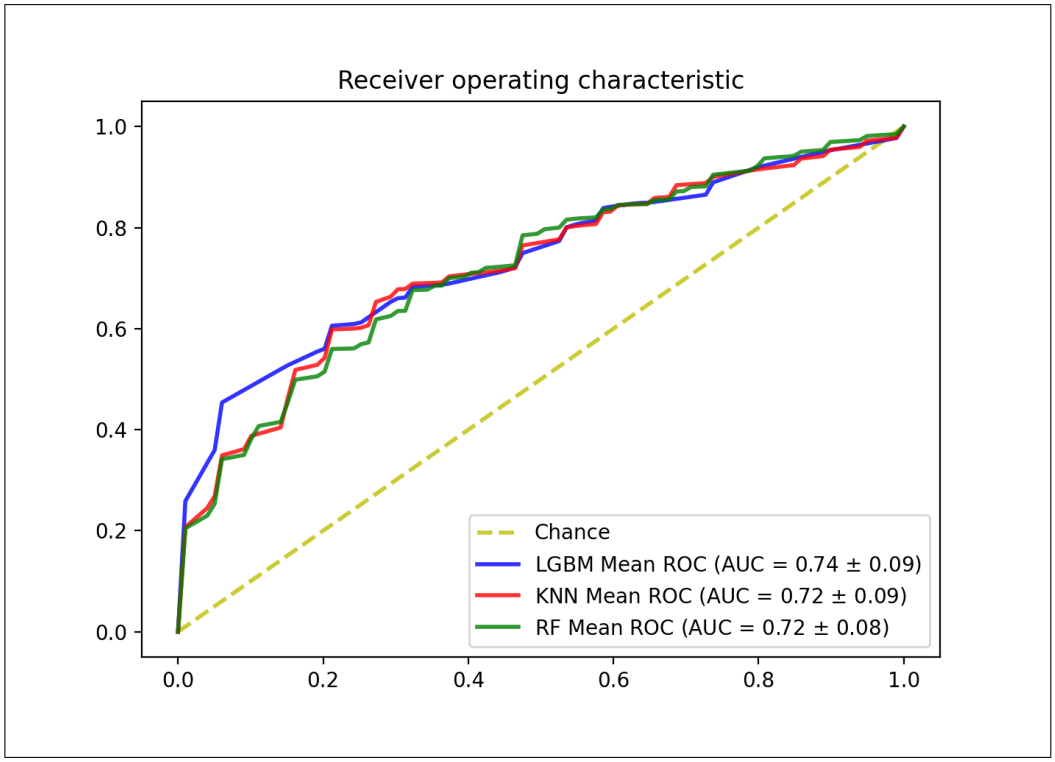

Figure A.17 Performances of three models trained with original features obtained from 2mm eTV.

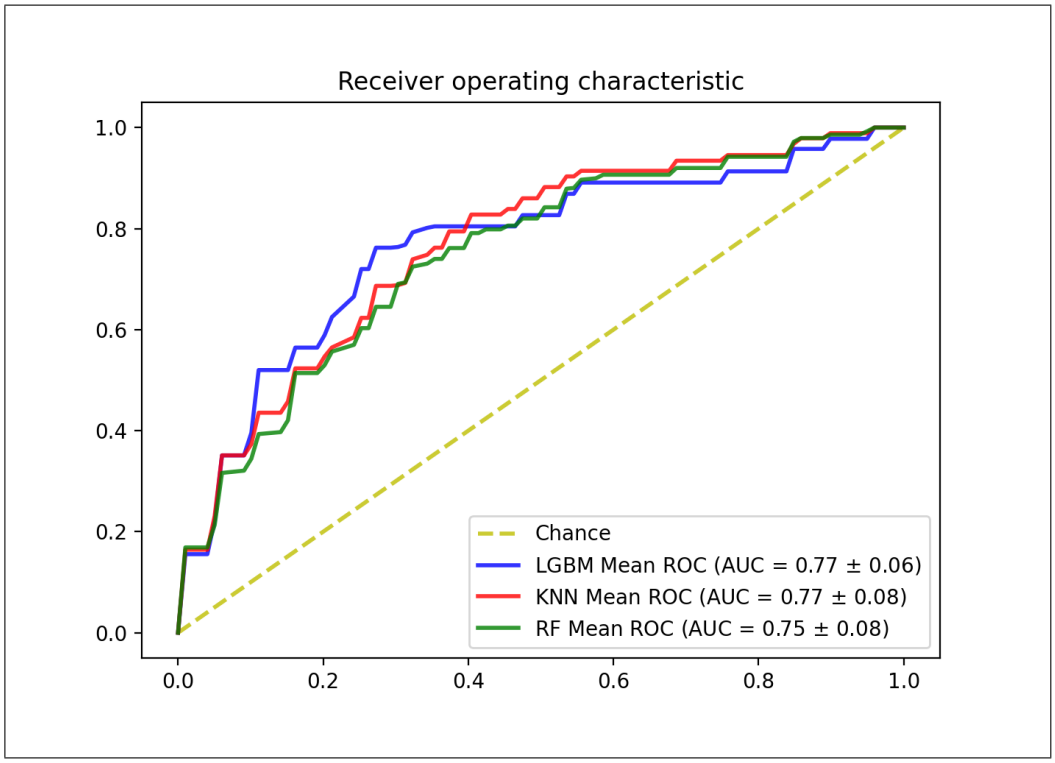

Figure A.18 Performances of three models trained with up-sampled (ADASYN) features obtained from 2mm eTV.

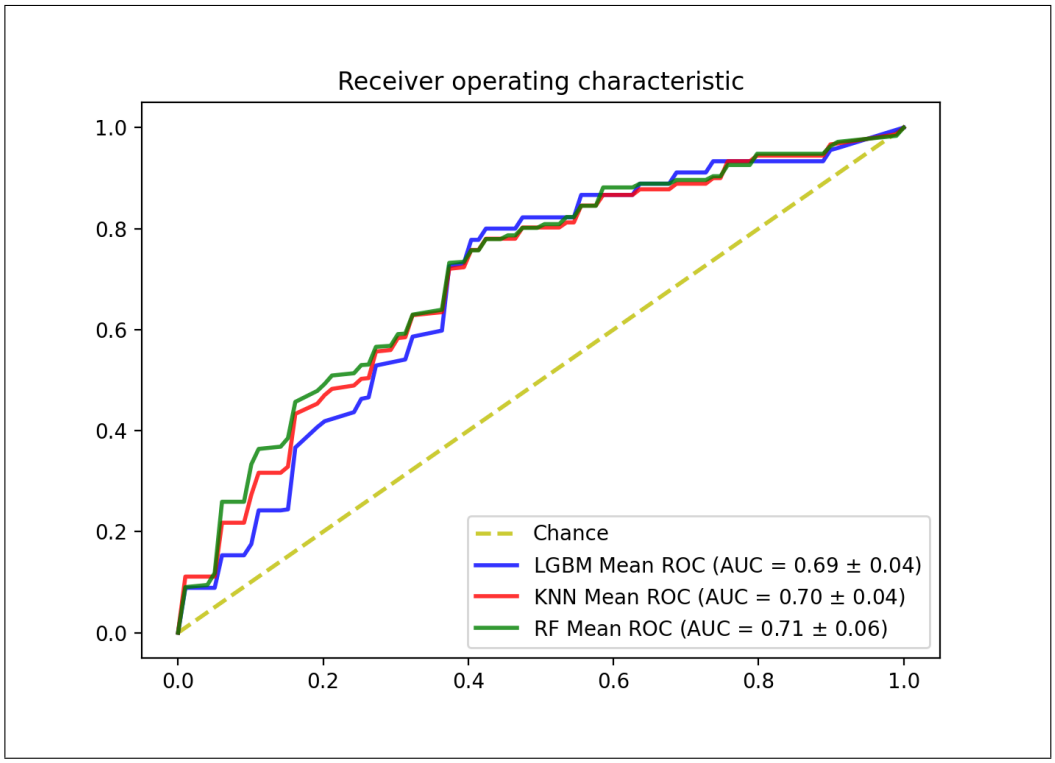

Figure A.19 Performances of three models trained with up-sampled (SMOTE) features obtained from 2mm eTV.

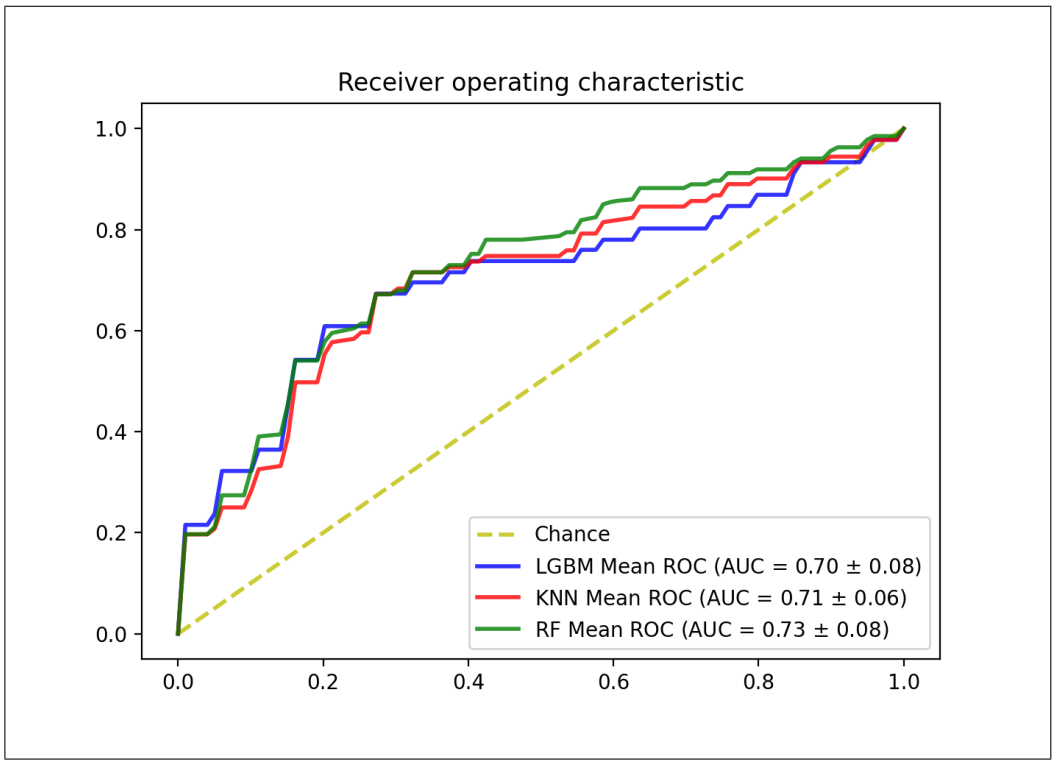

Figure A.20 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from 2mm eTV.

A.6 ROC Curves and Selected Features of 4mm eTV

Performance of three ML models trained with imbalance feature sets extracted from 4mm eTV are shown in Figure A.21. Performance of three ML models trained with using balanced/upsampled (ADASYN) feature sets extracted from 4mm eTV are shown in Figure A.22. Performance of three ML models trained with using balanced/upsampled (SMOTE) feature sets extracted from 4mm eTV are shown in Figure A.23. Performance of three ML models trained with using balanced/upsampled (SVMSMOTE) feature sets extracted from 4mm eTV are shown in Figure A.24. Mean ROC curves and AUC values of 5 fold nested cross validation are shown on images.

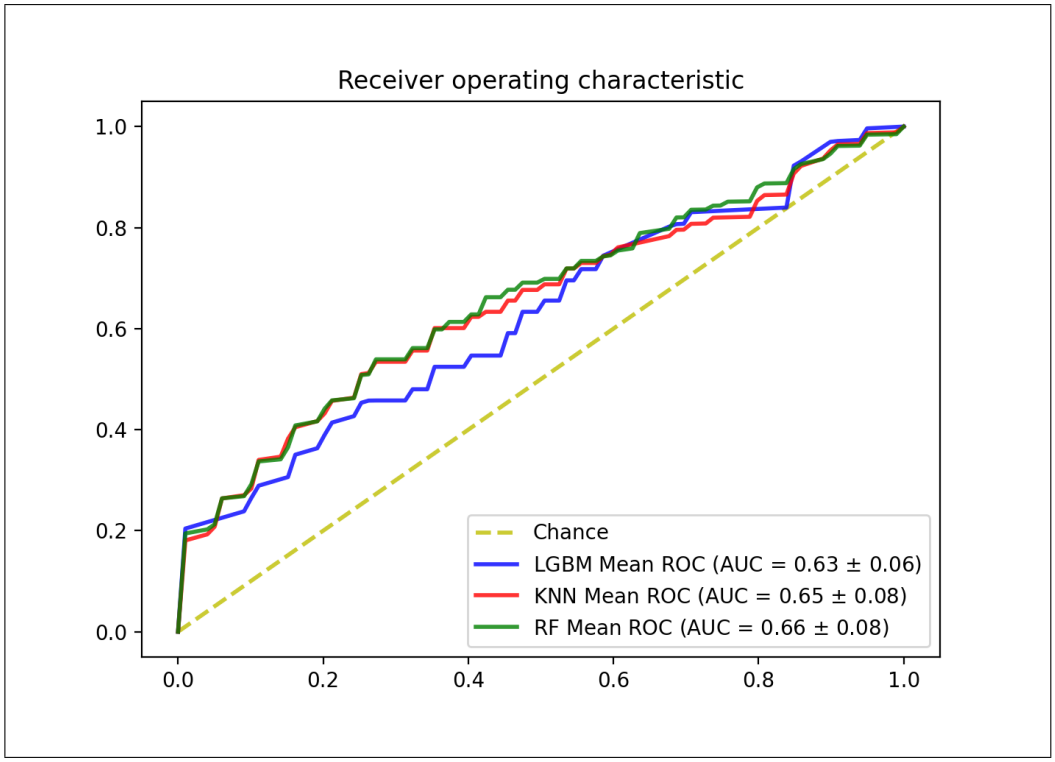

Figure A.21 Performances of three models trained with original features obtained from 4mm eTV.

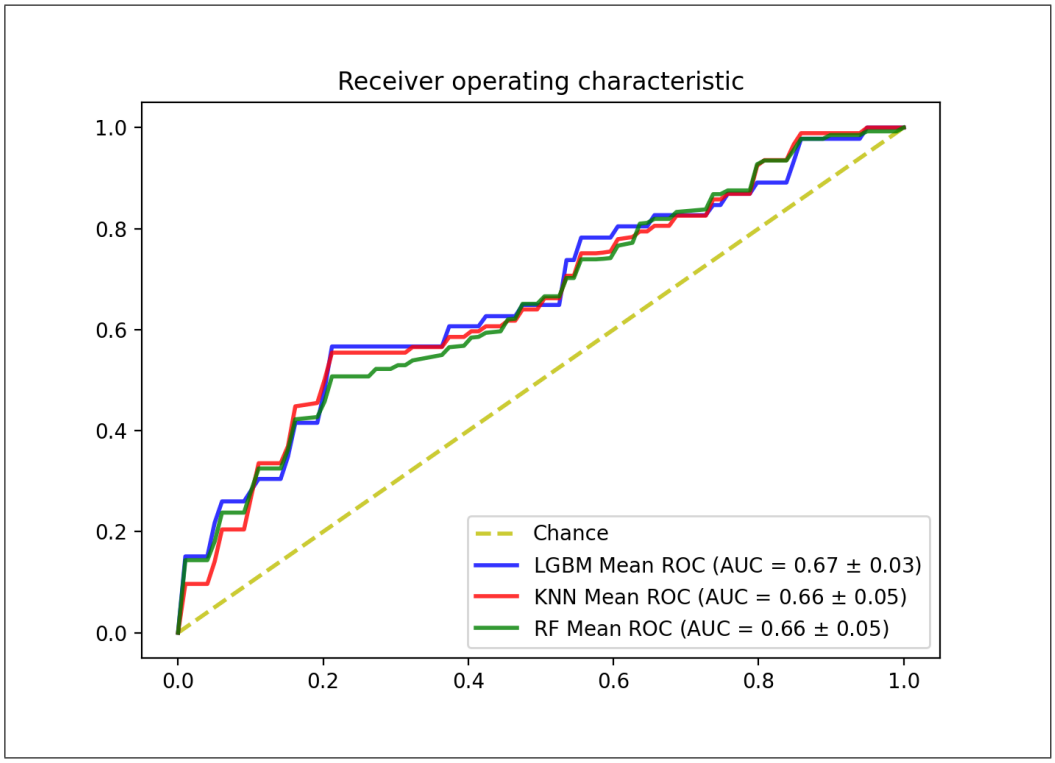

Figure A.22 Performances of three models trained with up-sampled (ADASYN) features obtained from 4mm eTV.

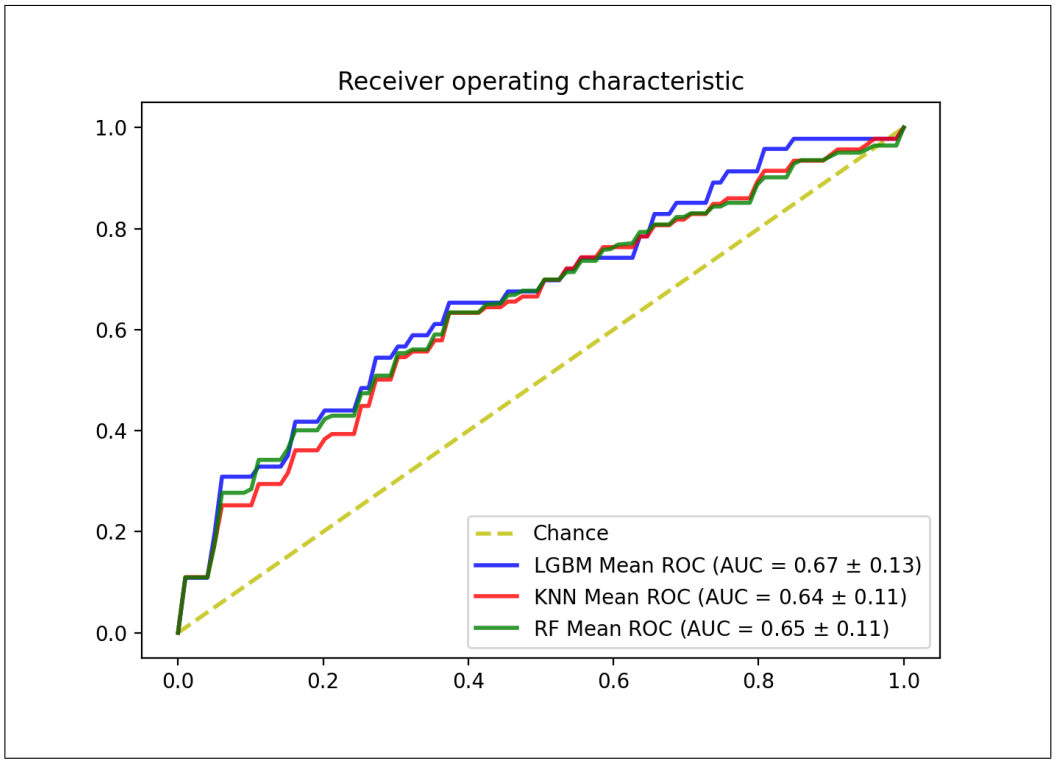

Figure A.23 Performances of three models trained with up-sampled (SMOTE) features obtained from 4mm eTV.

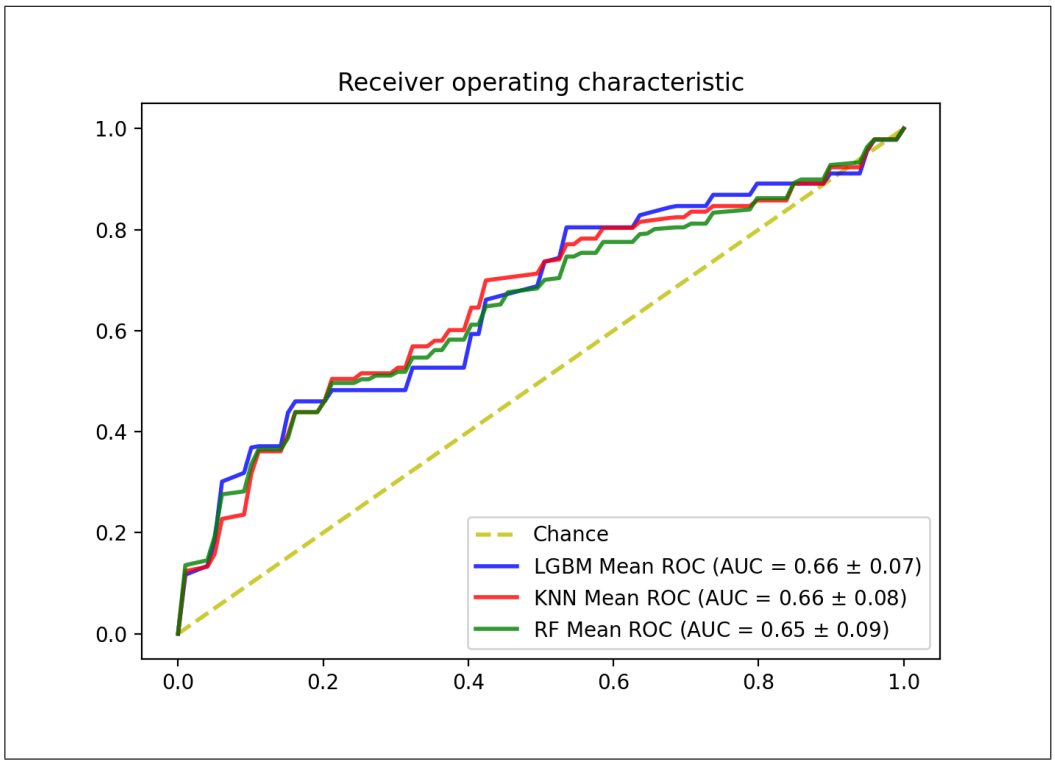

Figure A.24 Performances of three models trained with up-sampled (SVMSMOTE) features obtained from 4mm eTV.

REFERENCES

1. Sung, H., J. Ferlay, R. L. Siegel and M. Laversanne et al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, Vol. 71, pp. 209–249, June 2021.

2. Ferlay, J., M. Colombet, I. Soerjomataram and D. M. Parkin et al., "Cancer statistics for the year 2020: An overview," *International Journal of Cancer*, Vol. 149, pp. 778–789, August 2021.

3. Cairnis, P., "Renal cell carcinoma," *Cancer Biomarkers*, Vol. 9, pp. 461–473, October 2011.

4. Wan, Z., T. Yin, H. Chen, and D. Li, "Surgical treatment of a retroperitoneal benign tumor surrounding important blood vessels by fractionated resection: A case report and review of the literature," *Oncology Letters*, Vol. 11, pp. 3259–3264, May 2016.

5. Warren, A. Y., and D. Harrison, "Who/isup classifcation, grading and pathological staging of renal cell carcinoma: standards and controversies," *World Journal of Urology*, Vol. 36, pp. 1913–1926, 2018.

6. Moch, H., A. L. Cubilla, P. A. Humphrey and V. E. Reuter et al., "The 2016 who classification of tumours of the urinary system and male genital organs-part a: Renal, penile, and testicular tumours," *European Urology*, Vol. 70, no. 1, pp. 93–105, 2016.

7. Athanazio, D. A., L. S. Amorim, I. W. Cunha and K. R. M. Leite et al., "Classification of renal cell tumors - current concepts and use of ancillary tests: recommendations of the brazilian society of pathology," *Surgical and Experimental Pathology*, Vol. 4, February 2021.

8. Srigley, J. R., B. Delahunt, J. N. Eble and L. Egevad et al., "The international society of urological pathology (isup) vancouver classification of renal neoplasia," *The American Journal of Surgical Pathology*, Vol. 37, no. 10, pp. 1469–1489, 2013.

9. Muglia, V. F., and A. Prando, "Renal cell carcinoma: histological classification and correlation with imaging findings," *Radiol Bras*, Vol. 48, pp. 166–174, June 2015.

10. Vasudev, N. S., M. Wilson, G. D. Stewart and A. Adeyoju et al., "Challenges of early renal cancer detection: symptom patterns and incidental diagnosis rate in a multicentre prospective uk cohort of patients presenting with suspected renal cancer," *BMJ Open*, Vol. 10, May 2020.

11. Gray, R. E., and G. T. Harris, "Renal cell carcinoma: Diagnosis and management," *American Academy of Family Physicians*, Vol. 99, no. 3, pp. 179–184, 2019.

12. Gudbjartsson, T., T. J. Jónasdóttir, A. Thoroddsen and G. V. Einarsson et al., "A population-based familial aggregation analysis indicates genetic contribution in a majority of renal cell carcinomas," *International Journal of Cancer*, Vol. 100, pp. 476–479, August 2002.

13. Hemminki, K., and X. Li, "Familial risks of cancer as a guide to gene identification and mode of inheritance," *International Journal of Cancer*, Vol. 110, pp. 291–294, June 2004.

14. Mucci, L. A., J. B. Hjelmborg, J. R. Harris and K. Czene et al., "Familial risk and heritability of cancer among twins in nordic countries," *JAMA*, Vol. 315, pp. 68–76, January 2016.

15. Nguyen, K. A., J. S. Syed, C. R. Espenschied and H. LaDuca et al., "Advances in the diagnosis of hereditary kidney cancer: Initial results of a multigene panel test," *Wiley Online Library*, Vol. 123, pp. 4363–4371, November 2017.

16. Sharp, V. J., K. T. Barnes, and B. A. Erickson, "Assessment of asymptomatic microscopic hematuria in adults," *Am Fam Physician*, Vol. 88, pp. 747–754, December 2013.

17. Ljungberg, B., L. Albiges, Y. Abu-Ghanem and K. Bensalah et al., "European association of urology guidelines on renal cell carcinoma: The 2019 update," *European Urology*, Vol. 75, pp. 799–810, May 2019.

18. Grover, V. P. B., J. M. Tognarelli, M. M. E. Crossey and I. J. Cox et al., "Magnetic resonance imaging: Principles and techniques: Lessons for clinicians," *J Clin Exp Hepatol*, Vol. 5, pp. 246–255, September 2015.

19. Escudier, B., C. Porta, M. Schmidinger and N. Rioux-Leclercq et al., "Renal cell carcinoma: Esmo clinical practice guidelines for diagnosis, treatment and follow-up," *Annals of Oncology*, Vol. 27, pp. 58–68, September 2016.

20. Zhang, H., Q. Gan, Y. Wu and R. Liu et al., "Diagnostic performance of diffusion-weighted magnetic resonance imaging in differentiating human renal lesions (benignity or malignancy): a meta-analysis," *Abdominal Radiology*, Vol. 41, pp. 1997–2010, October 2016.

21. Vargas, H. A., J. Chaim, R. A. Lefkowitz and Y. Lekhman et al., "Renal cortical tumors: Use of multiphasic contrast-enhanced mr imaging to differentiate benign and malignant histologic subtypes," *Radiology*, Vol. 264, pp. 779–788, September 2012.

22. Jonisch, A. I., A. N. Rubinowitz, P. G. Mutalik, and G. M. Israel, "Can high-attenuation renal cysts be differentiated from renal cell carcinoma at unenhanced ct?," *Radiology*, Vol. 243, May 2007.

23. Pooler, B. D., P. J. Pickhardt, S. D. O. Conner and R. J. Bruce et al., "Renal cell carcinoma: Attenuation values on unenhanced ct," *American Journal of Roentgenology*, Vol. 198, pp. 1115–1120, May 2012.

24. Young, J. R., D. Margolis, S. Sauk and A. J. Pantuck et al., "Clear cell renal cell carcinoma: discrimination from other renal cell carcinoma subtypes and oncocytoma at multiphasic multidetector ct," *Radiology*, Vol. 267, pp. 444–453, May 2013.

25. Kopka, L., U. Fischer, G. Zoeller and C. Schmidt et al., "Dual-phase helical ct of the kidney: value of the corticomedullary and nephrographic phase for evaluation of renal lesions and preoperative staging of renal cell carcinoma," *AJR Am J Roentgenol*, Vol. 169, pp. 1573–1578, Dec 1997.

26. Dhaun, N., C. O. Bellamy, D. C. Cattran, and D. C. Kluth, "Utility of renal biopsy in the clinical management of renal disease," *International Society of Nephrology*, Vol. 85, pp. 1039–1048, May 2013.

27. Stoian, M., A. Dumitrache, and V. Stoica, "Past and present of renal biopsy in the management of patients with glomerular diseases," *Intemal MEdicine*, Vol. 16, pp. 31–40, October 2019.

28. Skinner, D. G., R. B. Colvin, C. D. Vermillion and R. C. Pfister et al., "Diagnosis and management of renal cell carcinoma. a clinical and pathologic study of 309 cases," *Cancer*, Vol. 28, pp. 1165–1177, Nov 1971.

29. Fuhrman, S. A., L. C. Lasky, and C. Limas, "Prognostic significance of morphologic parameters in renal cell carcinoma," *Am J Surg Pathol*, Vol. 6, pp. 655–663, October 1982.

30. Delahunt, B., J. C. Cheville, G. Martignoni and P. A. Humphrey et al., "The international society of urological pathology (isup) grading system for renal cell carcinoma and other prognostic parameters," *Am J Surg Pathol*, Vol. 37, pp. 1490–1504, October 2013.

31. Lambin, P., E. Rios-Velazquez, R. Leijenaar and S. Carvalho et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, Vol. 48, pp. 441–446, March 2012.

32. Lambin, P., R. T. H. Leijenaar, T. M. Deist and J. Peerlings et al., "Radiomics: the bridge between medical imaging and personalized medicine," *Nature Reviews Clinical Oncology*, Vol. 14, pp. 749–762, October 2017.

33. Strimbu, K., and J. A. Tavel, "What are biomarkers?," *Curr Opin HIV AIDS*, Vol. 5, pp. 463–466, Nov 2011.

34. Berenguer, R., M. del R. Pastor-Juan, J. Canales-Vázquez and M. Castro-Garcia et al., "Radiomics of ct features may be nonreproducible and redundant: Influence of ct acquisition parameters," *Radiology*, Vol. 288, April 2018.

35. Kumar, V., Y. Gu, S. Basu and A. Berglund et al., "Radiomics: the process and the challenges," *Magn Reson Imaging*, Vol. 30, pp. 1234–1248, Aug 2012.

36. Rizzo, S., F. Botta, S. Raimondi and D. Origgi et al., "Radiomics: the facts and the challenges of image analysis," *European Radiology Experimental volume*, Vol. 2, November 2018.

37. Limkin, E. J., R. Sun, L. Dercle and E. I. Zacharaki et al., "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology," *Ann Oncol*, Vol. 25, pp. 1191–1206, Jun 2017.

38. Torre, L. A., F. Bray, R. L. Siegel and J. Ferlay et al., "Global cancer statistics, 2012," *CA Cancer J Clin*, Vol. 65, pp. 87–108, Mar 2015.

39. Capitanio, U., and F. Montorsi, "Renal cancer," *Lancet*, Vol. 27, pp. 894–906, Feb 2016.

40. Novara, G., G. Martignoni, W. Artibani, and V. Ficarra, "Grading systems in renal cell carcinoma," *J Urol*, Vol. 177, pp. 430–436, Feb 2007.

41. Delahunt, B., J. N. Eble, L. Egevad, and H. Samaratunga, "Grading of renal cell carcinoma," *Histopathology*, Vol. 74, pp. 4–17, Jan 2019.

42. Lohse, C. M., M. L. Blute, H. Zincke and A. L. Weaver et al., "Comparison of standardized and nonstandardized nuclear grade of renal cell carcinoma to predict outcome among 2,042 patients," *Anatomic Pathology*, Vol. 118, pp. 877–886, 2002.

43. Bernhard, J. C., P. Bigot, G. Pignot and H. Baumert et al., "The accuracy of renal tumor biopsy: analysis from a national prospective study," *World J Urol*, Vol. 33, pp. 1205–1211, Aug 2015.

44. Gillies, R. J., P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, Vol. 278, pp. 563–577, Feb 2016.

45. Davnall, F., C. S. P. Yip, G. Ljungqvist and M. Selmi et al.,"Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?," *Insights Imaging*, Vol. 3, pp. 573–589, Dec 2012.

46. Ganeshan, B., and K. A. Miles, "Quantifying tumour heterogeneity with ct," *Cancer Imaging*, Vol. 13, pp. 140–179, Mar 2013.

47. Feng, Z., Q. Shen, Y. Li, and Z. Hu, "Ct texture analysis: a potential tool for predicting the fuhrman grade of clear-cell renal carcinoma," *Cancer Imaging*, Vol. 19, February 2019.

48. Bektas, C. T., B. Kocak, A. H. Yardimci and M. H. Turkcanoglu et al., "Clear cell renal cell carcinoma: Machine learning-based quantitative computed tomography texture analysis for prediction of fuhrman nuclear grade," *Eur Radiol*, Vol. 29, pp. 1153–1163, Mar 2019.

49. Ding, J., Z. Xing, Z. Jiang and J. Chen et al., "Ct-based radiomic model predicts high grade of clear cell renal cell carcinoma," *Eur Radiol*, Vol. 103, pp. 51–56, Jun 2018.

50. Zhou, Z., X. Qian, J. Hu and X. Ma et al., "Ct-based peritumoral radiomics signatures for malignancy grading of clear cell renal cell carcinoma," *Abdom Radiol (NY)*, Vol. 46, pp. 2690–2698, Jun 2021.

51. Cao, X. H., I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinformatics*, Vol. 17, Jun 2016.

52. Nguyen, H. M., E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, Vol. 3, pp. 4–21, 2011.

53. He, H., Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008.

54. Ke, G., Q. Meng, T. Finley and T. Wang et al., "Lightgbm: A highly efficient gradient boosting decision tree," *NIPS*, December 2017.

55. Breimen, L., "Random forests," *Machine Learning*, Vol. 45, pp. 5–32, December 2001.

56. Amin, A., D. Dori, P. Pudil, and H. Freeman, eds., *Nearest neighbors in random subspaces*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

57. Clark, K., B. Vendt, K. Smith and J. Freymann et al., "The cancer imaging archive (tcia): Maintaining and operating a public information repository," *J Digit Imaging*, Vol. 26, pp. 1045–1057, Jul 2013.

58. Heller, N., N. Sathianathen, A. Kalapara and E. Walczak et al., "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *J Digit Imaging*, 2019.

59. Shafiq-UI-Hassan, M., G. G. Zhang, K. Latifi and G. Ullah et al., "Intrinsic dependencies of ct radiomic features on voxel size and number of gray levels," *Med Phys*, Vol. 44, pp. 1050–1062, Mar 2017.

60. Yaniv, Z., B. C. Lowekamp, H. J. Johnson, and R. Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," *J Digit Imaging*, Vol. 31, pp. 290–303, Jun 2018.

61. Wang, Z., and K. Wang, *Cubic B-Spline Interpolation and Realization*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

62. van Griethuysen, J. M., A. Federov, C. Parmar and A. Hosny et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Res*, Vol. 77, Nov 2017.

63. Larue, R. T. H. M., J. E. van Timmeren, E. E. C. de Jon and, G. Feliciani et al., "Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thicknesses: a comprehensive phantom study," *Acta Oncol*, Vol. 56, pp. 1544–1553, Nov 2017.

64. Mukaka, M., "Statistics corner: A guide to appropriate use of correlation coefficient in medical research.," *Malawi medical journal : the journal of Medical Association of Malawi*, Vol. 24 3, pp. 69–71, 2012.

65. Guyon, I., and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, Vol. 3, pp. 1157–1182, mar 2003.

66. Reif, D. M., A. A. Motsinger, B. A. McKinney and J. E. Crowe et al. "Feature selection using a random forests classifier for the integrated analysis of multiple data types," in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pp. 1–8, IEEE, 2006.

67. Nogueira, S., K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *Journal of Machine Learning Research*, Vol. 18, no. 174, pp. 1–54, 2018.

68. DeLong, E., D. DeLong, and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, Vol. 44, pp. 837–845, September 1988.

69. Fluss, R., D. Faraggi, and B. Reiser, "Fluss r, faraggi d, reiser bestimation of the youden index and its associated cutoff point. biom j 47: 458-472," *Biometrical journal. Biometrische Zeitschrift*, Vol. 47, pp. 458–72, 08 2005.

70. Santos, M. S., J. P. Soares, P. H. Abreu and H. Araujo et al., "Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [research frontier]," *IEEE Computational Intelligence Magazine*, Vol. 13, no. 4, pp. 59–76, 2018.