Radiomics Machine Learning Analysis of Clear Cell Renal Cell Carcinoma for Tumour Grade Prediction Based on Intra-tumoural Subregion Heterogeneity

Abeer J. Alhussaini *, J. Douglas Steele , Adel Jawli , Ghulam Nabi *

Posted Date: 4 April 2024

doi: 10.20944/preprints202312.1379.v5

Keywords: clear cell renal cell carcinoma; renal masses; biopsy; computed tomography; radiomics; machine learning; tumour subregions; tumour heterogeneity; precision medicine

Article

Radiomics Machine Learning Analysis of Clear Cell Renal Cell Carcinoma for Tumour Grade Prediction Based on Intra-Tumoural Subregion Heterogeneity

Abeer J. Alhussaini [1,2,*], J. Douglas Steele [1], Adel Jawli [1,3] and Ghulam Nabi [1,*]

[1] Division of Imaging Sciences and Technology, School of Medicine, Ninewells Hospital, University of Dundee, Dundee DD1 9SY, UK
[2] Department of Clinical Radiology, Al-Amiri Hospital, Ministry of Health, Kuwait City, Kuwait
[3] Department of Clinical Radiology, Sheikh Jaber Al-Ahmad Al-Sabah Hospital, Ministry of Health, Kuwait City, Kuwait
* Correspondence: abr.hussaini@gmail.com, a.j.a.h.m.alhussaini@dundee.ac.uk (A.J.A.); g.nabi@dundee.ac.uk (G.N.)

Simple Summary: Clear cell renal cell carcinoma (ccRCC) accounts for at least 80% of renal tumours worldwide. The grading of clear cell carcinoma is crucial for its management; therefore, it is important to distinguish the ccRCC grade pre-operatively. The aim of this research is to differentiate high- from low-grade ccRCC non-invasively using machine learning (ML) and radiomics features extracted from pre-operative computed tomography (CT) scans, taking into consideration the tumour sub-region that offers the greatest accuracy when grading. Furthermore, radiomics and machine learning were compared with biopsy-determined grading in a sub-group with resection histopathology as a reference standard.

Abstract: *Background:* Renal cancers are among the top ten causes of cancer-specific mortality, of which the ccRCC subtype is responsible for most cases. The grading of ccRCC is important in determining tumour aggressiveness and clinical management. *Objectives:* The objectives of this research were to predict the WHO/ISUP grade of ccRCC pre-operatively and characterise the heterogeneity of tumour sub-regions using radiomics and ML models, including comparison with pre-operative biopsy-determined grading in a sub-group. *Methods:* Data were obtained from multiple institutions across two countries, including 391 patients with pathologically proven ccRCC. For analysis, the data were separated into four cohorts. Cohorts 1 and 2 included data from the respective institutions from the two countries, cohort 3 was the combined data from both cohort 1 and 2, and cohort 4 was a subset of cohort 1, for which both the biopsy and subsequent histology from resection (partial or total nephrectomy) were available. 3D image segmentation was carried out to derive a voxel of interest (VOI) mask. Radiomics features were then extracted from the contrast-enhanced images, and the data were normalised. The Pearson correlation coefficient and the XGBoost model were used to reduce the dimensionality of the features. Thereafter, 11 ML algorithms were implemented for the purpose of predicting the ccRCC grade and characterising the heterogeneity of sub-regions in the tumours. *Results:* For cohort 1, the 50% tumour core and 25% tumour periphery exhibited the best performance, with an average AUC of 77.9% and 78.6%, respectively. The 50% tumour core presented the highest performance in cohorts 2 and 3, with average AUC values of 87.6% and 76.9%, respectively. With the 25% periphery, cohort 4 showed AUC values of 95.0% and 80.0% for grade prediction when using internal and external validation, respectively, while biopsy histology had an AUC of 31.0% for the classification with the final grade of resection histology as a reference standard. The CatBoost classifier was the best for each of the four cohorts with an average AUC of 80.0%, 86.5%, 77.0% and 90.3% for cohorts 1, 2, 3 and 4 respectively. *Conclusion:* Radiomics signatures combined with ML have the potential to predict the WHO/ISUP grade of ccRCC with superior performance, when compared to pre-operative biopsy. Moreover, tumour sub-regions contain useful information that should be analysed independently when determining the tumour grade. Therefore, it is possible to distinguish the grade of ccRCC pre-operatively to improve patient care and management.

 doi:10.20944/preprints202312.1379.v5

Keywords: clear cell renal cell carcinoma; renal masses; biopsy; computed tomography; radiomics; machine learning; tumour subregions; tumour heterogeneity; precision medicine

1. Introduction

The grading of RCC has been acknowledged as a prognostic marker for close to a century [1]. The tumour grade provides some insight into how cancer may act. It identifies whether cancer cells are regular or aberrant under a microscope. The more aberrant the cells seem and the higher the grade, the quicker the tumour is likely to spread and expand. Many different grading schemes have been proposed, initially focused on a collection of cytological characteristics and more recently on nuclear morphology. Nuclear size (area, major axis, and perimeter), nuclear shape (shape factor and nuclear compactness), and nucleolar prominence characteristics are the main emphasis in the Fuhrman grading of renal cell carcinomas. Even though Fuhrman grading has been widely used in clinical investigations, its predictive value and reliability are up for discussion [2]. Fuhrman et al. [3] showed, in 1982, that tumours of grades 1, 2, 3, and 4 presented considerably differing metastatic rates. When grade 2 and 3 tumours were pooled into a single cohort, they likewise demonstrated a strong correlation between tumour grade and survival [3]. The International Society of Urologic Pathologists (ISUP) suggested a revised grading system for RCC in 2012, in order to address the shortcomings of the Fuhrman grading scheme [4]. This system is primarily based on the assessment of nucleoli and has been approved for papillary and ccRCC tumours [4]. The World Health Organization (WHO) recommended the ISUP grading system at a consensus conference in Zurich; as a result, the WHO/ISUP grading system is currently applied internationally [5].

Grading ultimately facilitates the optimal management and treatment of tumours, according to their prognostic behaviour concerning their respective grades. For instance, in elderly or very sick patients who have small renal tumours (<4 cm) and high mortality rate, cryoablation, active surveillance, or radiofrequency ablation may be considered to manage their conditions. It is crucial to note that confident radiological diagnosis of low-grade tumours in active surveillance can significantly impact clinical decisions, hence eliminating the risk of over-treatment. As ccRCC is the most prevalent subtype (8 in 10 RCCs) with the highest potential for metastasis, it requires careful characterisation [6]. High-grade cancers have a poorer prognosis, are more aggressive, have high risk of post-operative recurrence, and may metastasise. Therefore, it is very important to differentiate between different grades of ccRCC, as high-grade ccRCC requires immediate and exact management. Precision medicine together with personalised treatment has advanced with the advent of cutting-edge technology; hence, clinicians are interested in determining the grade of ccRCC before surgery or treatment, enabling them to better advise regarding therapy and even predict cancer-free survival if surgery has been conducted.

The diagnosis of ccRCC grade is commonly carried out based on pre- and post-operative methods. One such pre-operative method is biopsy. However, the accuracy of a biopsy can be influenced by several factors, including the size and location of the tumour, the experience of the pathologist performing the biopsy, and the quality of the biopsy sample [7]. Due to sampling errors, a biopsy may not always provide an accurate representation of the overall tumour grade [8]. Inter-observer variability can also lead to inconsistencies in the grading process. This can be especially problematic for tumours that are borderline between two grades [9]. In some cases, a biopsy may not provide a definitive diagnosis as it only considers the cross-sectional area of the tumour and, compounded with the fact that ccRCC presents high spatial and temporal heterogeneity [10], it may not be representative of the entire tumour [11]. A biopsy also has a small chance of haemorrhage (3.5%) and a rare risk of track seeding (1:10,000) [12]. Due to the limitations highlighted for biopsies [13], radical or partial nephrectomy treatment specimens are usually used as definitive post-operative diagnostic tools for tumour grade. With partial or radical nephrectomy being the definitive therapeutic approach, a small but significant number of patients are subjected to unnecessary surgery, even though their management

 doi:10.20944/preprints202312.1379.v5

may not require surgical resection. Nephrectomy also increases the possibility of contracting chronic renal diseases that may result in cardiovascular ailments. Therefore, precise grading of ccRCC through non-invasive methods is imperative, in order to improve the effectiveness and targeted management of tumours.

The assumption in most research and clinical practice is that solid renal masses are homogenous in nature or, if heterogeneous, they have the same distribution throughout the tumour volume [14]. Recent studies [15] have highlighted that, in some histopathologic classifications, different tumour sub-regions may have different rates of aggressiveness; hence, heterogeneity plays a significant role in tumour progression. Ignoring such intra-tumoural differences may lead to inaccurate diagnosis, treatment, and prognoses. The biological makeup of tumours is complex and, therefore, leads to spatial differences within their structures. These variations may be due to the expression of genes or the microscopic structure [16]. Such differences can be caused by several factors, including hypoxia (i.e., the loss of oxygen in the cells) and necrosis (i.e., the death of cells). This is mostly synonymous with the tumour core. Likewise, high cell growth and tumour-infiltrating cells are factors associated with the periphery [17].

Medical imaging analysis has been shown to be capable of detecting and quantifying the level of heterogeneity in tumours [18–20]. This ability enables tumours to be categorised into different sub-regions depending on the level of heterogeneity. In relation to tumour grading, intra-tumoural heterogeneity may prove useful in determining the sub-region of the tumour containing the most prominent features that enable successful grading of the tumour. Radiomics, which is the extraction of high-throughput features from medical images, is a modern technique that has been used in medicine to extract features that would not be otherwise visible to the naked eye alone [21]. It was first proposed by Lambin et al. [22] in 2012 to extract features, taking into consideration the differences in solid masses. Radiomics eliminates the subjectivity in the extraction of tumour features from medical images, functioning as an objective virtual biopsy. A significant number of studies have applied radiomics approaches for the classification of tumour subtypes, grading, and even the staging of tumours [23,24].

Artificial intelligence (AI) is a wide area whose aim is to build machines which simulate human cognitive abilities. It has enabled a shift from human systems to machine systems trained by computers using features obtained from the input data. In recent years, with the advent of AI, there has been tremendous progress in the field of medical imaging. Machine learning which is a branch of AI has been used to extract high dimension features from medical images and have shown significant ability to perform image segmentation, recognition, reconstruction and classification. They have also made it easy to quantify and standardise medical image features thereby acting as an intermediary between clinics and pathology. AI has been proven to be effective in reducing misdiagnosis and improving accuracy in the diagnosis of renal diseases. ML is expected by numerous researchers to bring drastic changes in the field of individualised diagnosis and patient treatment and is currently used to predict the nuclear grade, classification and prognosis of RCC using radiomic data [25]. AI has also enabled the analysis of tumour sub-regions in a variety of clinical tasks, using several imaging modalities such as CT and MRI [26]. However, these analyses have been limited to only a few types of tumours, particularly brain tumours [27], head and neck tumours [28], and breast cancers [29]. To date, no study has attempted to analyse the effect of subregion intra-tumoural heterogeneity on the diagnosis, treatment, and prognosis of renal masses and specifically ccRCCs. This rationale formed the basis of the present study, focusing on the effect of subregional intra-tumoural heterogeneity on the grading of ccRCC. To the best of our knowledge, this is the first paper to comprehensively focus on tumour sub-regions in renal tumours for the prediction of tumour grade.

In this research, the hypothesis that radiomics combined with ML can significantly differentiate between high- and low-grade ccRCC for individual patients is tested. This study sought solutions to two major problems that previous research has not been able to address:

- *Characterising the effect of subregional intra-tumoural heterogeneity on grading in ccRCC;*

 doi:10.20944/preprints202312.1379.v5

- *Comparing the diagnostic accuracy of radiomics and ML through image-guided biopsy in determining the grade of renal masses, using resection (partial or complete) histopathology as a reference standard.*

1.1. Key Contributions:

- ***Clinical Application:*** *This research offers a practical application of radiomics and machine learning techniques in the field of oncology, specifically in the diagnosis and grading ccRCC which could potentially aid clinicians in early detection, accurate and precise treatment planning due to its higher diagnostic accuracy in comparison to traditional methods.*
- ***Subregion Heterogeneity Analysis:*** *The inclusion of intra-tumoural subregion heterogeneity analysis highlights the depth of the research beyond simple tumour delineation and delves into the spatial distribution and variations within the tumour. This provides deeper insights into tumour biology and behaviour, leading to more personalised treatment approaches.*
- ***Potential for Non-invasive Assessment:*** *Radiomics–machine learning algorithms, has the potential to extract valuable information from medical images non-invasively reducing the need for invasive procedures for tumour characterisation and grading, improving patient comfort and reducing healthcare costs.*

2. Materials and Methods

2.1. Ethical Approval

This study was approved by the institutional board, and access to patient data was granted under the Caldicott Approval Number IGTCAL11334 dated 21 October 2022. Informed consent for the research was not required, as CT scan image acquisition is a routine examination procedure for patients suspected of having ccRCC.

2.2. Study Cohorts

This retrospective multi-centre study used data from three centres either in partnership with or satellite hospitals of the National Health Service (NHS) in a well-defined geographical area of Scotland, United Kingdom. The institutions included Ninewells Hospital Dundee, Stracathro General Hospital, and Perth Royal Infirmary Hospital. Data from the University of Minnesota Hospital and Clinic (UMHC) was also used [30,31]. Scan data was anonymised.

We accessed the Tayside Urological Cancers (TUCAN) database for pathologically confirmed cases of ccRCC between January 2004 and December 2022. A total of 396 patients with CT scan images were retrieved from the Picture Archiving and Communication System (PACS) in DICOM format. This data formed our first cohort (cohort 1). Retrospective-based analysis for pathologically confirmed ccRCC image data following partial or radical nephrectomy from UMHC stored in a public database [32] (accessed on 21 May 2022) was performed, referred to as cohort 2 in this study. The database was queried for data between 2010 and 2018. Data for a total of 204 patients with ccRCC CT scan images was collected.

2.2.1. Inclusion Criteria

- Availability of protocol-based pre-operative contrast-enhanced CT scan in the arterial phase. The selection of the arterial phase is justified due to its widespread adoption across medical centres. Moreover, this phase, characterised by its enhancement pattern and hypervascularity, has demonstrated promising capabilities in distinguishing between low- and high-grade ccRCC [33].
- Confirmed histopathology from partial or radical nephrectomy with grades reported by a uro-pathologist according to the WHO/ISUP grading system.
- CT scans with data to achieve a working acquisition for 3D image reconstruction.

2.2.2. Exclusion Criteria

- Patients with only biopsy histopathology.

 doi:10.20944/preprints202312.1379.v5

- Metastatic clear cell renal cell carcinoma (mccRCC).
- Patients with bilateral or ipsilateral multiple tumours, primarily due to the ambiguity of the database in distinguishing between the exact tumour grades.

For more information on the UHMC data set, the reader is referred to [30,31].

2.3. CT Acquisition Technique

The patients in cohort 1 were examined using up to five different CT helical/spiral scanners, including GE Medical Systems, Philips, Siemens, Canon Medical Systems, and Toshiba 512-row detectors. The detectors were also of different models, including Aquilion, Biograph128, Aquilion Lightning, Revolution EVO, Discovery MI, Ingenuity CT, LightSpeed VCT, Brilliance 64, Aquilion PRIME, Aquilion Prime SP, and Brilliance 16P. The slice thicknesses were 1.50, 0.63, 2.00, 1.25, and 1.00 mm. Likewise, the number of pixels in the image was 512x512. The arterial phase of the CT scan obtained 20–45 seconds after contrast injection was acquired using the following method: intravenous Omnipaque 300 contrast agent (80–100 ml/s), 3 ml/s contrast injection for the renal scan, and 100–120 kVp with an X-ray tube current of 100–560 mA depending on the size of the patient. For the UMHC data set, refer to [30,31].

2.4. Hardware and Software Consideration

A windows 10 machine was used with the following hardware: Device name: ASUS FX503VM, Processor: Intel(R) Core(TM) i7–7700HQ CPU @ 2.80 GHz, Nvidia GeForce GTX 1060 3GB GPU, CUDA Cores: 1152, Base Clock: 1506MHz, Boost Clock: 1708MHz, Texture Units: 72, Memory Clock: 8GHz, Memory Bandwidth: 192GB/s, ROPs: 48.L2, Cache Size: 1536KB, Installed RAM: 32.0 GB (2*16 GB) DD–2400MHz and System type: 64–bit operating system, x64–based processor.

2.5. Data Curation

The procedure used for data collection with respect to each patient comprised multiple stages: accessing the Tayside Urological Cancers database, identification of patients using a unique identifier (community health index number (CHI) number), review of the medical records of the cohort, CT data acquisition, annotation of the data and, finally, quality assurance. Anonymised data for cohort 1 were in DICOM format. For each patient, duplicated DICOM slices were removed, as slice inconsistencies has detrimental effects on how an image can be processed.

Image quality is an important factor in machine learning modelling [34]. Therefore, as usual practice [33,35,36] qualitative measures were used to remove images of low quality. An expert diagnostic imaging technologist (A.J.A.) visually inspected each of the images and further verified by the co-authors. Images with severe blurring, granularity (quantum mottle), and ring and metal artefacts [37] were removed. Figure 1 presents a flowchart showing the exclusion and inclusion criteria for patients and their sample size distribution. Tumour grades 1 and 2 were labelled as low-grade, whereas grades 3 and 4 were classified as high-grade. This is due to the clinical management for grades 1 and 2 being more or less similar, which is also the case for grades 3 and 4.

 doi:10.20944/preprints202312.1379.v5

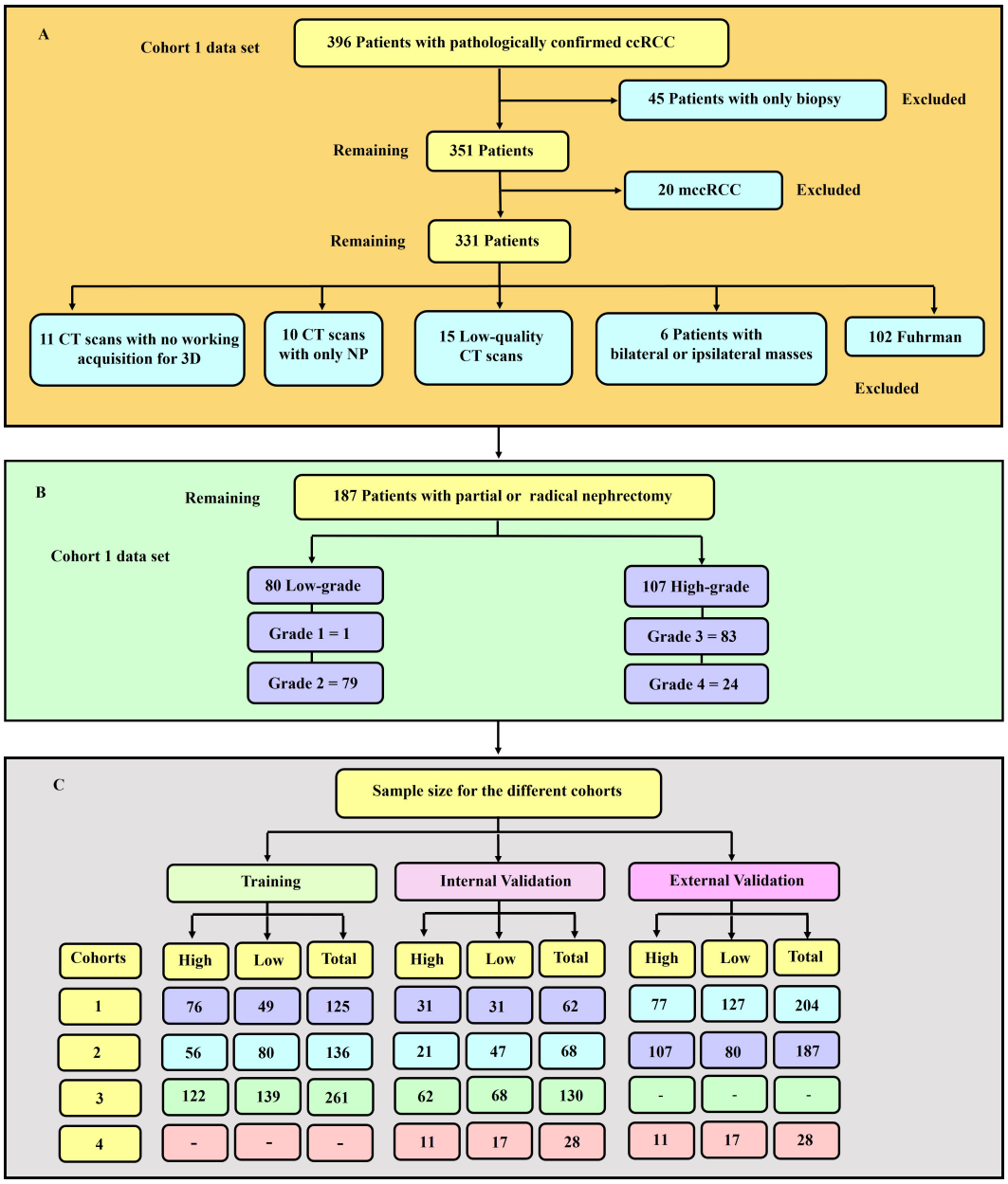


Figure 1. Diagrammatic representation of the exclusion and inclusion criteria for the cohort 1 data set and samples distribution for different cohorts.

2.6. Tumour Sub-Volume Segmentation Technique

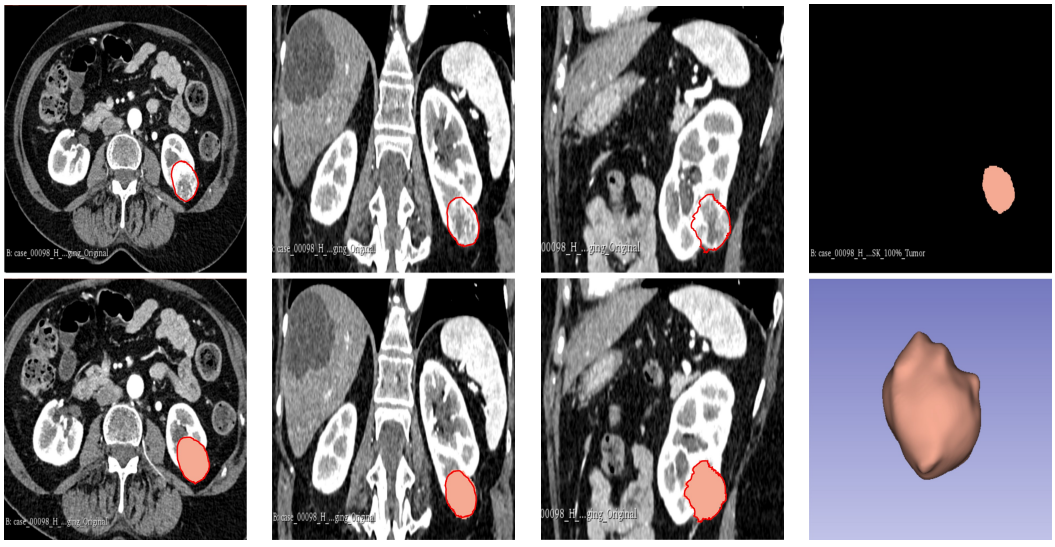
In cohort 1, CT image slices for each patient were converted to 3D NIFTI (Neuroimaging Informatics Technology Initiative) format using the Python programming language version 3.9, then loaded into the 3D Slicer version 4.11.20210226 software for segmentation. Manual segmentation was performed on the 3D images, delineating the edges of the tumour slice-by-slice to obtain the VOI.

The above procedure was performed by a blinded investigator (A.J.A.) with 14 years of experience in interpreting medical images, who was unaware of the final pathological grade of the tumour. Confirmatory segmentation was carried out by another blinded investigator (A.J.) with 12 years of experience in using medical imaging technology on 20% of the samples, in order to ascertain the accuracy of the first segmentation. Thereafter, the segmentations were assessed and ascertained by an independent experienced urological surgical oncologist (G.N.), taking into consideration radiology and histology reports. The gold standard pathology diagnosis was assumed to be partial or radical nephrectomy histopathology.

 doi:10.20944/preprints202312.1379.v5

For cohort 2, Heller et al. [30] previously conducted segmentation by following a set of instructions, including ensuring that the images of the patients contain the entire kidney, drawing a contour which includes the entire tumour capsule and any tumour or cyst but excluding all tissues other than renal parenchyma, and drawing a contour that includes the tumour components but excludes all kidney tissues. In the present study, only the delineation of kidney tumours achieved by Heller et al. [30] was considered. To perform delineation, a web-based interface was created on the HTML5 Canvas, which allowed contours to be drawn freehand on the images. The image series were sub-sampled in the longitudinal direction regularly, such that the number of annotated slices depicting a kidney was about 50% that of the original. Interpolation was also performed. More information on the segmentation of the cohort 2 data set can be found in the report [30].

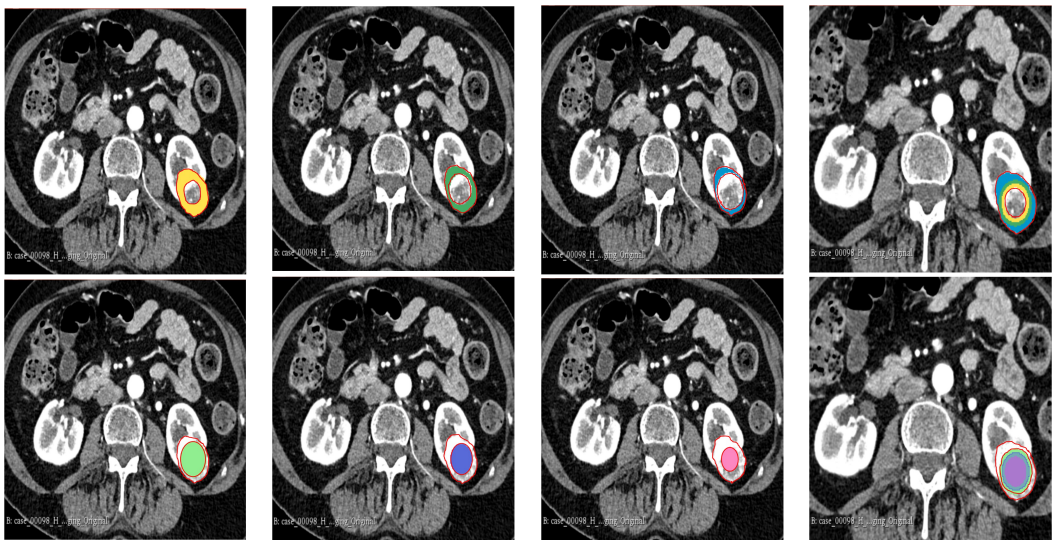
The segmentation result for both cohorts 1 and 2 was a binary mask of the tumour. In the present study, the tumours were divided into different sub-regions based on their geometry (i.e., periphery and core). The periphery refers to regions towards the edges of the tumour, whereas the core represents regions close to the centre of the tumour. The core was obtained by extracting 25%, 50%, and 75% of the binary mask from the centre of the tumour, while the periphery was generated 'by extracting 25%, 50%, and 75% of the binary mask starting from the edges of the tumour to form a rim as a hollow sphere. A detailed visual description is shown in Figure 2, Figure 3, and Figure 4. Mask generation was performed using a Python script which automatically generated the sub-regions by image subtraction techniques.



(**a**) (**b**) (**c**) (**d**)

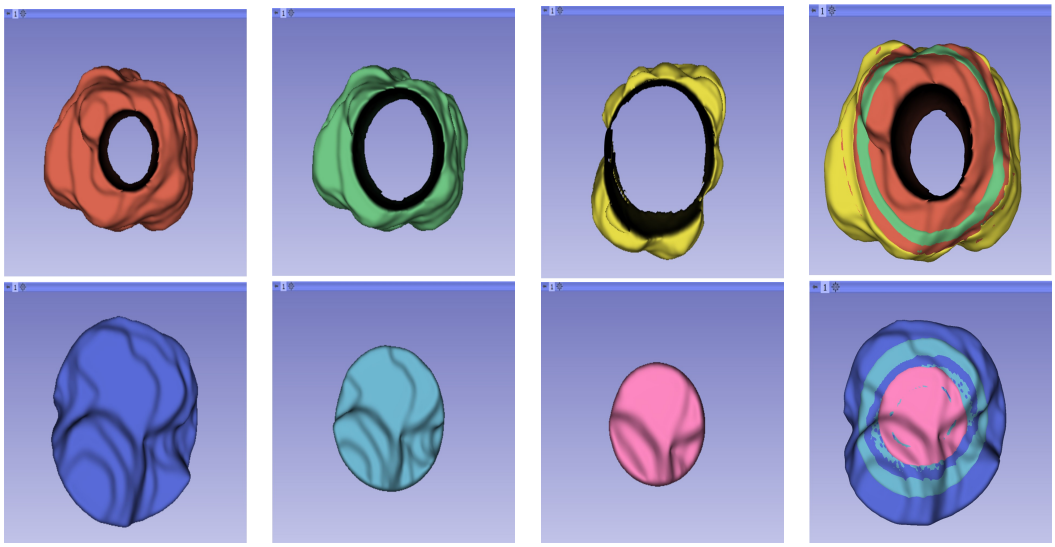
Figure 2. Manual segmentation of the 3D image slices using the 3D Slicer software: (**a**) Axial plane; (**b**) Coronal plane; (**c**); Sagittal plane; and (**d**) Generated 3D VOI from the delineated 2D slices.

 doi:10.20944/preprints202312.1379.v5



(**a**) (**b**) (**c**) (**d**)

Figure 3. Manual segmentation of the 3D image slices using the 3D Slicer software. (**a**) 75% periphery and core of the tumour; (**b**) 50% periphery and core of the tumour; (**c**) 25% periphery and core of the tumour; and (**d**) Overlap of periphery and core sub-regions.



(**a**) (**b**) (**c**) (**d**)

Figure 4. Representation of the 3D segmented regions: (**a**) 75% periphery and core of the tumour; (**b**) 50% periphery and core of the tumour; (**c**) 25% periphery and core of the tumour; and (**d**) Overlap of periphery and core sub-regions.

2.7. Radiomics Feature Computation

Similar to our previous research [23], texture descriptors of the features were computed using the PyRadiomics module in Python 3.6.1. The aim of the PyRadiomics module is to implement a standardised method for extracting radiomic features from medical images, thus avoiding inter-observer variability [38]. The parameters used in PyRadiomics were a minimum region of interest (ROI) dimension of 2, a pad distance of 5, a normalisation value of false, and normaliser scale of 1. There was no removal of outliers, no re-sampled pixel spacing, and no pre-cropping of the image. SitkBSpline was used as the interpolator, with the bin width set to 20.

 doi:10.20944/preprints202312.1379.v5

On average, PyRadiomics generated approximately 1500 features for each image and enabled the extraction of 7 feature classes per 3D image. The extracted feature categories were as follows: First-order (19 features), grey-level co-occurrence matrix (GLCM) (24 features), grey-level run-length matrix (GLRLM) (16 features), grey-level size-zone matrix (GLSZM) (16 features), grey-level dependence matrix (GLDM) (14 features), neighbouring grey-tone difference matrix (NGTDM) (5 features), and 3D shape features (16 features). These features allow for computation of texture intensities as well as their distribution in the image [38].

In a previous study [23], it has been shown that a combination of the original feature classes and filter features significantly improved the model performance. Therefore, we extracted the filter class features in addition to the original features. These filter classes included the local binary pattern (LBP-3D), gradient, exponential, logarithm, square-root, square, Laplacian of Gaussian (LoG), and wavelet. The filter features were applied to every feature in the original feature classes; for instance, as the first-order statistic feature class has 19 features, it follows that it will have 19 LBP filter features. The filter class features were named according to the name of the original feature and the name of the filter class. [38].

2.8. Feature Processing and Feature Selection

The features extracted using PyRadiomics were standardised to assume a standard distribution. Scaling was performed using the Z-score, as shown in Equation (1), for both the training and testing data sets independently, using only the mean and standard deviation generated from the training set. This was carried out to avoid leakage of information while also eliminating bias from the model. All of the features were transformed in such a way that they followed a standard normal distribution with mean (μ) = 0 and standard deviation (σ) = 1.

$$Z = (x - \mu)/\sigma, \tag{1}$$

where

Z: Value after scaling the feature.
x: The feature.
μ: Mean of all features in the training set.
σ: Standard deviation of the training set.

Normalisation reduces the effect of different scanners, as well as any influence that intra-scanner differences may introduce in textural features, resulting in improved correlation to histopathological grade [39]. The ground-truth labels were denoted as 1 for high-grade and 0 for low-grade, for the purpose of enabling the ML models to understand the data. Machine learning models usually encounter the "curse of dimensionality" in the training data set [40], when the number of features in the data set is greater than the number of samples. Therefore, we applied two feature selection techniques in an attempt to reduce the number of features and retain only those features with the highest importance in predicting the tumour grade. First, the inter-feature correlation coefficients were computed and, when two features had a correlation coefficient greater than 0.8, one of the features was dropped. Thereafter, we used the XGBoost algorithm to further select the features with the highest importance for model development.

2.9. Sub-Sampling

In ML, the distribution of data among different classes is an important consideration before developing an ML model. An imbalance in the data may cause model bias; instead of learning the features of the data, “cramming” occurs, making the model inapplicable in real-life scenarios. In this research, our data samples were imbalanced; therefore, we applied the synthetic minority oversampling technique (SMOTE). Care should be taken when using SMOTE, as it should only be

 doi:10.20944/preprints202312.1379.v5

applied in the training set and not the validation or testing sets; if this occurs, then there is a possibility that the model gains an unrealistic improvement in operational performance due to data leakage [41].

2.10. Statistical Analysis

Common clinical features in this research were analysed using the SciPy package. Comparisons were made based on age, gender, tumour size, and tumour volume against the pathological grade. The chi-squared test (χ^2) was conducted to compare the associations between categorical groups. It is a non-parametric test that is used when the data do not follow the assumptions of parametric tests, such as the assumption of normality in the distribution of the data. The Student's t-test is a popular statistical tool, which is used when assessing the difference between two population means for continuous data; it is normally used when the population follows a normal distribution and the population variance is unknown. The point-biserial correlation coefficient (rpb) was calculated to further confirm the significance in cases where statistical significance between clinical data was obtained. The Pearson correlation coefficient (r) was used to measure the linear correlations in the data between the radiomic features. The value of this coefficient ranges between -1 and $+1$, with $+1$ signifying a strong positive correlation.

McNemar's statistical test, which is a modified chi-squared test, was used to test whether the difference between false negative (FN) and false positive (FP) was statistically significant. It was calculated from the confusion matrix using the stats module in the SciPy library. The chi-squared test for randomness was used to test whether the model predictions differed from random predictions. The Dice similarity coefficient was used to determine the inter-reader agreement for the segmentations. All statistical tests assumed a significance level of $p < 0.05$ (i.e., the null hypothesis was rejected when the p-value was less than 0.05). The radiomic quality score (RQS) was also calculated, in order to evaluate whether the research followed the scientific guidelines of radiomic studies. The study followed the established guidelines of transparent reporting of a multi-variable prediction model for individual prognosis or diagnosis (TRIPOD) [42].

2.11. Model Construction, Validation, and Evaluation

Several models were implemented to predict the pathological grade of ccRCC, using the WHO/ISUP grading system as the gold standard. The choice of the models implemented was motivated by a previous research [23,24,33,36,43–50] where the models provided satisfactory results for tumour subtype prediction using radiomics. The models were constructed for cohorts 1, 2, and the combined cohort. The ML models included the support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost/XGB), naïve Bayes (NB), multi-layer perceptron classifier (MLP), long short-term memory (LSTM), logistic regression (LR), quadratic discriminant analysis (QDA), light gradient boosting machine (LightGBM/LGB), category boosting (CatBoost/CB), and adaptive boosting (AdaBoost/ADB). Different parameters were tested for each model to arrive at the optimum parameters giving the best results. Refer to Table 1 for the parameter optimisation of the models.

In total, 231 distinct models were constructed: 11 models for each of the 3 cohorts and each tumour sub-region ($11 \times 3 \times 7$). For validation, the data set was divided into training and testing sets. For the three main cohorts 1,2 and 3, 67% of the data were used for training and 33% were retained for testing. This formed part of our internal validation procedure. Moreover, cohort 1 was validated against cohort 2 and vice versa, forming part of our external validation procedure. It should be noted that, although cohort 1 was taken to be analogous to a "single institution" data set, it was obtained from multiple centres, and its comparison with cohort 2 was for the purpose of external validation of the predictive models.

A subset of cohort 1 consisting of patients who underwent both a CT-guided percutaneous biopsy and nephrectomy (28 samples) was evaluated using two separate ML algorithms. One model was trained on cohort 1 but excluding the 28 patients who had both procedures conducted, while the second model was trained on cohort 2, acting as an external validator. The classifiers and sub-regions were

 doi:10.20944/preprints202312.1379.v5

determined for the two best-performing classifiers and the three best-performing tumour regions. The objective of this test was to assess the accuracy of tumour grade classification in biopsy histopathology when compared to ML prediction using partial or total nephrectomy histopathology as the gold standard. These 28 samples are referred to as cohort 4. In situations where the biopsy grade histopathology was indeterminate for a specific tumour, we concluded its final pathological grade as the opposite of the nephrectomy grade for that tumour (i.e., if the nephrectomy outcome was high-grade but the biopsy result was indeterminate, we concluded that the biopsy was low-grade for the purpose of analysis). It is worth-noting that "Indeterminate results" are important as they do not contribute to decision-making and one of the reasons biopsy approaches has not been adopted amongst clinician's world over. In fact, our group addressed this issue by 3x2 tables in a systematic review published earlier [51]. We believe that all studies should report "indeterminate results" for the sake of transparency and external validity.

Evaluation of the model performance was carried out using a number of metrics, including accuracy (ACC), specificity (SPE), sensitivity (SEN), area under the curve of the receiver operating characteristic curve (AUC-ROC), the Matthews correlation coefficient (MCC), F1 score, McNemar's test (McN), and the chi-squared test (χ^2). All major metrics were reported at 95% Confidence Interval (CI).

In a highly imbalance data set, accuracy is not reliable as it gives an overly optimistic measure of the majority class [52]. MCC is an effective solution to overcome class imbalance and has been applied by the Food and Drug Administration (FDA) agency in the USA for the evaluation of MicroArrayII/Sequencing Quality control (MAQC/SEQC) projects [53]. There are situations however where MCC give undefined or large fluctuations in their results [54]. The combination of precision/recall which is the F1-score [55] provide better information than the pair of sensitivity/specificity [56] and has gained popularity since the 1990's in the machine learning world. Despite its popularity, F1-score varies for class swapping while MCC is invariant. The AUC-ROC curve is a popular evaluation metric used when a single threshold of the confusion matrix is unavailable [57]. It is also sensitive to class imbalance though it is widely used in the medical field therefore it is used in this study to compare with previous research. All the other metrics have been reported as well and it is at the readers discretion to compare which metrics suits the study.

Table 1. Parameter optimisation for machine learning models.

MODELS	PARAMETERS
SVM	kernel=rbf, probability=True, random_state=42, gamma=0.2, C=0.01
RF	n_estimators=401, random_state=42, max_depth=3
XGBoost	random_state=42, learning_rate=0.01, n_estimators=401, gamma=0.52
NB	GaussianNB
MLP	hidden_layer_sizes=(401,201), activation=relu, solver=adam, max_iter=5
LSTM	loss=binary_crossentropy, optimizer=Adam, lr=0.01, metrics=accuracy, epochs=1000, batch_size=16
LR	random_state=42, max_iter=4
QDA	reg_param=0.05
LightGBM	random_state=42, n_estimators=9
CatBoost	random_state=42, verbose=False, iterations=50
AdaBoost	base_estimator=rf, n_estimators=201, learning_rate=0.01, random_state=42

3. Results

3.1. Study Population and Statistical Analysis

In cohort 1, after implementation of the inclusion and exclusion criteria, a total of 187 patients with pathologically proven ccRCC were obtained. Of these, 80 patients presented low-grade and 107 presented high-grade ccRCC. The mean age was 59.05 and 64 years for low- and high-grade tumours,

 doi:10.20944/preprints202312.1379.v5

respectively. Gender-wise, 65.78% of patients were male and 34.22% were female. The average tumour size and tumour volume were 4.32 cm and 75.8 cm^3, respectively, for low-grade patients, and 6.033 cm and 203.74 cm^3, respectively, for high-grade patients. For cohort 2, the data set met all the inclusion and none of the exclusion criteria; hence, no sample was eliminated. The mean age was 57.17 and 63.68 years for low- and high-grade patients. The average tumour size and tumour volume were 3.89 cm and 51.44 cm^3 for low-grade patients, and 6.81 cm and 235.26 cm^3 for high-grade patients, respectively. In terms of gender, 65.69% of patients were male, while the rest were female.

Differences in average age, tumour size, and tumour volume (but not gender) were statistically significant in cohorts 1, 2, and the combined cohort. However, using the point-biserial correlation coefficient (rpb) to compare the correlation between the best model prediction and the clinical features, no statistically significant difference was found. Table 2 provides the characteristics and analysis of patients. The Dice similarity coefficient score was 0.93, which indicated that there was a good inter-reader agreement for tumour segmentation. The entire data set RQS was found to be 61.11%, signifying that the research followed scientific radiomic guidelines. For the RQS rubric, we refer the reader to https://www.radiomics.world/rqs2 [42].

Table 2. Statistical demographic characteristics of patient data.

Tumour and Patient Characteristics					
	Variable	**Low-Grade**	**High-Grade**	**p-Value**	**rpb*[1]**
Cohort 1 n=187	**Age (Mean ± SD)**	59.05±12.28	64±9.40	0.002*	0.4
	Size (cm)	4.32±2.02	6.03±3.23	0.00006*	0.6
	Volume (cm^3)	75.8±90.90	203.74±305.82	0*	0.6
	Gender			0.331	
	Male	49 (26.20%)	74 (39.57%)		
	Female	31 (16.58%)	33 (17.65%)		
Cohort 2 n=204	**Age (Mean ± SD)**	57.17±12.67	63.68±11.14	0*	0.88
	Size (cm)	3.89±2.16	6.81±3.55	0*	0.45
	Volume (cm^3)	51.44±114.32	235.26±326.10	0*	0.56
	Gender			0.25	
	Male	77 (37.75%)	57 (27.94%)		
	Female	50 (24.51%)	20 (9.80%)		
Cohort 3 n=391	**Age (Mean ± SD)**	57.89±12.55	63.86±10.17	0*	0.56
	Size (cm)	3.89±2.16	6.81±3.55	0*	0.2
	Volume (cm^3)	60.86±106.55	216.93±314.85	0*	0.29
	Gender			0.25	
	Male	126 (32.23%)	131 (33.50%)		
	Female	81 (20.72%)	53 (13.55%)		
Cohort 4 n=28	**Age (Mean ± SD)**	57.12±10.25	62.09±9.39	0.22	
	Size (cm)	3.31±0.94	4.02±2.25	0.28	
	Volume (cm^3)	25.98±27.10	57.16±70.40	0.13	
	Gender			1	
	Male	12 (42.86%)	8 (28.57%)		
	Female	5 (17.86%)	3 (10.71%)		

* Statistical significance at the 0.05 level; *[1] point-biserial correlation coefficient (rpb).

3.2. Feature Extraction, Pre-processing, and Selection

A total of 1875 features were extracted using the PyRadiomics library. It should be noted that there were no null values in the data, as it is crucial in the context of ML to handle these values to avoid errors and undefined results. The Pearson correlation coefficient (r) and extreme gradient boosting algorithm were used to reduce the number of features to an optimal number. The number of features selected (FS) varied between the data sets.

 doi:10.20944/preprints202312.1379.v5

3.2.1. Model Validation and Evaluation

Internal Validation

Cohort 1

Of the developed models, the CatBoost classifier performed the best for the majority of the tumour sub-region models, with its best classifier having an AUC of 85.0% in the 100% tumour. When the tumour sub-region was considered, the 50% tumour core and 25% tumour periphery exhibited the best performance, with an average AUC of 77.9% and 78.6%, respectively. When the models' core and periphery regions were averaged, the best classifier was CatBoost, with an AUC of 80.0% = (80.7%+79.3%)/2. Table A1 and Table A2 provide the results obtained for cohort 1.

Cohort 2

The best-performing model in the cohort 2 data set was the CatBoost classifier, with the best performance in the 50% tumour periphery having an AUC of 91.0%. In terms of tumour sub-region, the 50% tumour core had the highest average AUC of 87.6%. When the models' subregions were averaged, the best classifier was CatBoost, with an AUC of 86.5% = (87.0%+86.0%)/2. Table A3 and Table A4 provide the results obtained for cohort 2.

Cohort 3

When NHS and UMHC data were combined, the models with the highest AUC were the 50% tumour core CatBoost classifier and the 75% tumour periphery RF classifier, with AUC of 80.0% for both. The 50% tumour core was the best region, with an average AUC of 76.9%. When the models' core and periphery regions were averaged, the best classifiers were RF and CatBoost, with AUC values of 77.3% = (76.3%+78.3%)/2 and 77.0% = (77.3%+76.7%)/2, respectively. Table A5 and Table A6 provide the results obtained for cohort 3.

External Validation

Cohort 1

When cohort 2 was used as the training set and cohort 1 was predicted on its models, the best-performing model was the QDA 25% tumour periphery classifier, with an AUC of 71.0%. For the tumour sub-region, the 25% tumour periphery was the best, with an average AUC of 65.0%. When the models were averaged, the best classifier was QDA, with an AUC of 67.7% = (67.3%+68.0%)/2. Table A7 and Table A8 provide the relevant results.

Cohort 2

With cohort 1 as the training set and cohort 2 as the testing set, the best-performing model was the SVM 50% tumour core classifier, with an AUC of 77.0%. In terms of tumour sub-region, the 50% tumour core was the best, with an average AUC of 74.2%. When the models were averaged, the best classifier was RF, with an AUC of 74.8% = (74.3%+75.3%)/2. Refer to Table A9 and Table A10 for the results.

3.2.2. Comparison between Biopsy and Machine Learning Classification

When the biopsy classification results on the 28 samples separated from the NHS data set were compared to ML predictions, machine learning exhibited the highest AUC values of 95.0% and 80.0% for internal and external validation, respectively, using the CatBoost classifier. This was better than the AUC of 31.0% obtained from biopsy results, in terms of correctly grading renal cancer, as shown in Figure A1, Figure A2 and Figure A3. Relevant statistics are provided in Table 3 and Table A11.

 doi:10.20944/preprints202312.1379.v5

Table 3. Comparison of the best diagnostic performance in cohort 4 under biopsy and machine learning models.

	BIOPSY	MACHINE LEARNING	
METRICS		**INTERNAL VALIDATION**	**EXTERNAL VALIDATION**
ACC	35.7±18	96.4±5	82.1±14
SPE	52.9±24	100.0±12	88.2±14
SEN	9.1±13	90.9±13	72.7±26
AUC	**31.0±17**	**95.0±8**	**80.0±15**
MCC	-0.40	0.93	0.62
F1	0.1	0.95	0.76
McN	0.64	0.32	0.65
χ^2	0.15	0	0.06

4. Discussion

Clear cell renal cell carcinoma is the most common subtype of renal cell carcinoma and is responsible for the majority of renal cancer-related deaths. It comprises up to 80% of RCC diagnoses [58], and is more likely to metastasise to other organs. Important diagnostic criteria that must be derived include tumour grade, tumour stage, and the histological type of the tumour. For most cancer patients, histological grade is a crucial predictor of local invasion or systemic metastases, which may affect how well they respond to treatment. To define the extent of the tumour, tumour staging-based clinical assessment, imaging investigations, and histological assessment are required. A greater comprehension of the neoplastic situation and awareness of the limitations of diagnostic techniques are made possible through an understanding of the procedures involved in tumour diagnosis, grading, and staging.

To accurately grade a tumour, several grading schemas have been applied, of which the WHO/ISUP and Fuhrman grading systems are the most popular and widely accepted. Previously, grading was focused on a collection of cytological characteristics of the tumour; however, nuclear morphology has more recently become a major area of focus. The Fuhrman grading system has been used for some time [59], with its worldwide adoption in 1982 [3].

There are several methodological issues with the study conducted by Fuhrman et al.; for example, its reliance on retrospective data collected over a 13-year period raises questions about potential biases [1]. The system's dependency on a small sample size of only 85 cases may also make its conclusions less generalisable [1,59]. The inclusion of several RCC subtypes without subtype-specific grading eliminated the possibility of variations in tumour behaviour [1,59,60]. It is difficult to grade consistently and accurately, due to the complexity of the criteria, which call for the simultaneous evaluation of three nuclear factors (i.e., nuclear size, nuclear irregularity, and nucleolar prominence) [59,60], resulting in poor inter-observer reproducibility and interpretability. The lack of guidelines that can be utilised to assign weights to the different discordant parameters to achieve a final grade makes the Fuhrman system even more controversial [59,61]. Furthermore, the shape of the nucleus has not been well-defined for different grades [59]. Grading discrepancies are a result of conflict between the grading criteria and a lack of direction for resolving them [1,4,60]. Additionally, imprecise standards for nuclear pleomorphism and nucleolar prominence adversely affect classifications made by pathologists, resulting in increased variability [59]. Even if a tumour is localised, grading according to the highest-grade area could result in an over-estimation of tumour aggressiveness [1,59]. This system's inconsistent behaviour and poor reproducibility [60] have raised questions regarding its dependability and potential effects on patient care and prognosis [62]. Flaws regarding inter-observer repeatability [62], and the fact that the Fuhrman grading system is still widely used despite these flaws, indicate that there is a need for more research and better grading methods.

An extensive and co-operative effort resulted in the development of the ISUP grading system for renal cell neoplasia in 2012 as an alternative to the Fuhrman grading system [59,60]. The system was

 doi:10.20944/preprints202312.1379.v5

ratified and adopted by the WHO in 2015 and renamed as the WHO/ISUP grading system [1,5]. As opposed to the Fuhrman grading system, the ISUP system focuses on the prominence of nuclei as the sole parameter that should be utilised when identifying the tumour grade. This reduction in rating parameters has led to better grade distinction and increased predictive value. This has also eliminated the controversy around reproducibility that had been identified with respect to the Fuhrman grading system. Previous studies have shown that there is a clear separation between grades 2 and 3 in the WHO/ISUP grading system, which was not the case with the Fuhrman system. Indeed, in their study, Dugher et al. [4] have highlighted the downgrade of Fuhrman grades 2 and 3 to grades 1 and 2, respectively, in the WHO/ISUP system. This indicates that, besides the overlap of grades in the Fuhrman system, there was also an over-estimation of grades—a problem that has been rectified with the WHO/ISUP grading system [4,24]. The WHO/ISUP grading system has been highly associated with the prognoses of patients.

Pre-operative image-guided biopsy is a diagnostic tool that is used to identify the tumour grade. However, there are inherent problems that have been identified in connection with this approach, including the fact that it is invasive in nature, causes discomfort, and may lead to other complications in patients when the procedure is performed [13]. Therefore, non-invasive testing, imaging, and clinical evaluations may be necessary to confirm the presence of ccRCC and its grade without having to undergo such a procedure. Radiomics has gained traction in clinical practice in recent years, and has been a buzzword since 2016 [21]. It refers to the extraction of high-dimensional quantitative image features in what is known as image texture analysis, describing the pixel intensity in medical images such as X-ray images as well as CT, MRI, CT/PET, CT/SPECT, US, and mammogram scans. Radiomics approaches have been applied in a number of studies for the diagnosis, grading, and staging of tumours. Machine learning is one of the major branches of AI, providing methods that are trained on a set of known data and then tested on unknown data. In this way, researchers have attempted to make machines more intelligent through determining spatial differences in data that would have been otherwise difficult for a human being to decipher. Such approaches have been used in combination with texture analyses, particularly for tumour classification, grading, and staging. They are capable of learning and improving through the analysis of image textural features, thereby resulting in higher accuracy than native methods [63].

Heterogeneity within tumours is a significant predictor of outcomes, with greater diversity within the tumour being potentially linked to increased tumour severity. The level of tumour heterogeneity can be represented through images known as kinetic maps, which are simply signal intensity time curves [64–66]. Previous studies [67,68] that have utilised these maps typically end up averaging the signal intensity features throughout the solid mass; hence, regions with different levels of aggressiveness end up contributing equally to determining the final features. This leads to a loss of information regarding the correct representation of the tumour [69,70]. In some studies, there have been attempts to preserve intra-tumoural heterogeneity through extracting the features at the periphery and the core and analysing them separately [18–20,68]; however, this is still not sufficient, as information from other sub-regions of the tumour is not considered.

The objective of this work was to study the impact of subregion intra-tumoural heterogeneity on the grading of ccRCC, comparing and contrasting ML/AI based methods combined with CT radiomics signatures with biopsy and partial and radical nephrectomy in terms of determining the grade of ccRCC. Finally, the possibility of using CT radiomics ML analysis as an alternative to—and, thereby, as a replacement for—the conventional WHO/ISUP grading system in the grading of ccRCC was investigated.

The experimental findings of our research highlighted various aspects for discussion. From the results, it was found that age, tumour size, and tumour volume were statistically significant for cohorts 1, 2, and 3. However, for cohort 4, none of the clinical features were found to be significant. Upon further analysis of the statistically significant clinical features using the point-biserial correlation coefficient (rpb), no features were verified as significant. Furthermore, the 50% tumor core was

 doi:10.20944/preprints202312.1379.v5

identified as the optimal tumor sub-region, exhibiting the highest average performance across models in cohorts 1, 2, and 3, with average AUCs of 77.9%, 87.6%, and 76.9%, respectively. It is worth noting that the 25% tumour periphery presented an increase in average performance for cohort 1, having an AUC of 78.6%; however, this result was not statistically different from that of the 50% core, and it failed to register the best performance in the other cohorts.

Among the 11 classifiers, the CatBoost classifier was the best model in all three cohorts, with average AUC values of 80.0%, 86.5%, and 77.0% for cohorts 1, 2, and 3, respectively. Likewise, the best-performing distinct classifier per cohort was CatBoost, with AUC of 85.0% in the 100% core, 91.0% in the 50% periphery, and 80.0% in the 50% core for cohorts 1, 2, and 3, respectively. In the external validation, cohort 1 validated on cohort 2 data had the highest performance in the 25% periphery, with the highest AUC of 71.0% and the best classifier being QDA. Conversely, cohort 2 validated on cohort 1 data provided the best performance in the 50% core, with an AUC of 77.0% and the best classifier being the SVM. Finally, in the comparison between biopsy- and ML-based classification of the 28 patients who underwent both biopsy and nephrectomy (i.e., cohort 4), the ML model was found to be more accurate, with the best AUC values for internal and external validation being 95.0% and 80.0%, respectively, in comparison to an AUC of 31.0% when biopsy was performed. In this case, the nephrectomy results of grading were assumed as the ground-truth.

For each of the 231 models the pathological grade of a tumour was predicted in less than 2 seconds. It is worth noting that the segmentations in cohort 1 were markedly different from those in cohort 2. Cohort 2 emerged as the highest-performing group, followed by cohort 1, while the combined cohort, notably cohort 3, exhibited the lowest performance. This disparity can be attributed to several factors, including variations in scanners, segmentations, pixel size, section thickness, tube current, tube voltage, kernel reconstruction, enhancement of contrast agent and imaging protocols. Moreover, cohort 1, in itself is a multi-institutional data set from three different centres. This may contrebuted to the low performance during external validation. Refer to Table A1—Table A10 for comparison.

Clinical feature significance is an important aspect of any research, as it gives a general overview of the data to be used in a study. Few studies have opted to include clinical features which are statistically significant into their ML radiomics models [71,72]. Takahashi et al. [72], for instance, incorporated 9 out of 12 clinical features into their prediction model due to them being statistically significant [72]. In our study, age, tumour size, and tumour volume were found to be statistically significant; however, they were not integrated into the ML radiomics model as a confirmatory test using the point-biserial correlation coefficient revealed a lack of significance. Nonetheless, there is a lack of clear guidelines on the relationship between statistical significance and predictive significance. There is a misunderstanding that association statistics may result in predictive utility; however, association only provides information regarding a population, whereas predictive research focuses on either multi-class or binary classification of a singular subject [73]. Moreover, the degree of association between clinical features and the outcome is affected by sample size; that is, statistical significance is likely to increase with an increase in sample size [74]. This has been clearly portrayed in previous research, such as that of Alhussaini et al. [23]. Even in our own research, for the cohort 4 data—despite being derived from the same population as cohort 1—the age, tumour size, tumour volume, and gender were not statistically significant, indicating that the sample size might be the likely cause.

4.1. Literature Related to Methodological Proposal

Zhao et al. [44], in their prospective research, presented interesting findings regarding tumour sub-regions in ccRCC. In their research, they indicated that somatic copy number alterations (CANs), grade, and necrosis are higher in the tumour core, compared to the tumour margin. Our findings, obtained using different tumour sub-regions, tend to agree with the study by Zhao et al. [44], even though the authors did not construct a predictive ML algorithm. He et al. [45] constructed five predictive CT scan models using an artificial neural network algorithm, in order to predict the tumour grade of ccRCC using both conventional image features and texture features. The best-performing

 doi:10.20944/preprints202312.1379.v5

model in their study, using the corticomedullary phase (CMP) and the texture features, provided an accuracy of 91.8%. This is comparable to our study, in which the CatBoost classifier attained the highest accuracy of 91.1%. However, He et al. [45] did not use other metrics, which could have been useful in analysing the overall success of the prediction. For instance, the research could have depicted a high accuracy but with bias towards one class. Moreover, the research findings were not externally validated; hence, its performance is unclear with respect to other data sets.

Similar to He et al. [45], Sun et al. [36] constructed an SVM algorithm to predict the pathological grade of ccRCC. The results of their research gave an AUC of 87.0%, sensitivity of 83.0%, and specificity of 67.0%. However, we found that they erred by giving an overly optimistic AUC with a very low specificity. This can easily be seen by analysing our SVM results for the best-performing SVM model, which had an AUC of 86.0%, sensitivity of 81.0%, and specificity of 91.5%. Our best model—the CatBoost classifier—performed much better. Xv et al. [46] set out to analyse the performance of the SVM classifier using three feature selection algorithms for the differentiation of ccRCC pathological grades in both clinical–radiological and radiomics features. The three algorithms were the LASSO, recursive feature elimination (RFE), and ReliefF algorithms. Their best-performing model was SVM–ReliefF with combined clinical and radiomics features, with an AUC of 88.0% in the training set, 85.0% in the validation set, and 82.0% in the test set. It is worth noting that we used none of the feature selection algorithms used by Xv et al. [46], but obtained better performance.

Cui et al. [47] conducted internal and external validation for the purpose of predicting the pathological grade of ccRCC. Their research achieved satisfactory performance, with internal and external validation accuracy of 78.0% and 61.0%, respectively, in the corticomedullary phase using the CatBoost classifier. Compared to their research, our findings indicated better performance when the CatBoost classifier was used for both the internal and external validation, with an accuracy of 91.2% and 76.0%, respectively, in the CMP. Wang et al. [48] also conducted a multi-centre study using a logistic regression model; however, they used both biopsy and nephrectomy as the ground-truth, despite the challenges that have been highlighted regarding biopsies. They did not report on the internal validation performance; however, their training AUC, sensitivity, and specificity were 89.0%, 85.0%, and 84.0%, respectively. Likewise, their external validation AUC, sensitivity, and specificity were 81.0%, 58.0%, and 95.0%, respectively. Their external validation performance was better than our performance using the LR model, which obtained an AUC, sensitivity, and specificity of 74.0%, 59.7%, and 88.2%, respectively. However, in general, our CatBoost classifier still out-performed their LR model. Moldovanu et al. [49] investigated the use of multi-phase CT using LR to predict the WHO/ISUP nuclear grade of ccRCC. When our results were compared with their validation set, which yielded an AUC, sensitivity, and specificity of 81.0%, 72.7%, and 75.9% in the corticomedullary phase, our research exhibited higher performance not only in the best-performing model but also in the LR model, which obtained an AUC, sensitivity, and specificity of 84.0%, 71.4%, and 95.8%, respectively.

Yi et al. [50] have performed research for prediction of the WHO/ISUP pathological grade of ccRCC using both radiomics and clinical features with an SVM model. The 264 samples used were from the nephrographic phase (NP). We noted that there was a massive class imbalance in the data, with a ratio between low- and high-grade samples of 78:22; however, they did not highlight how this issue was resolved. Nonetheless, the testing accuracy yielded an AUC of 80.2%, lower than that obtained in our research. Similar to our study, Karagöz and Guvenis [43] constructed a 3D radiomic feature-based classifier to determine the nuclear grade of ccRCC using the WHO/ISUP grading system. The best results were obtained using the LightGBM model, which obtained an AUC of 0.89. They also carried out tumour dilation and contraction by 2 mm, which led them to conclude that the ML algorithm is robust against deviation in segmentation by observers. Our best model out-performed their research and our sample size was much larger, thereby providing more trustworthy results. Demirjian et al. [24] also constructed a 3D model using data from two institutions using RF, AdaBoost, and ElasticNet classifiers. The best-performing model, RF, obtained an AUC of 0.73. This model performance was lower than in our research. The use of a data set graded using the Fuhrman system for testing may

 doi:10.20944/preprints202312.1379.v5

have led to poor results, as WHO/ISUP and Fuhrman use different parameters for grading; hence, it is not advisable to use the Fuhrman grade as the ground-truth for a model trained using WHO/ISUP. Shu et al. [33] extracted radiomics features from the CMP and NP to construct 7 ML algorithms, with the best model in the CMP (i.e., the MLP algorithm) achieving an accuracy of 0.97. The findings of this study are quite interesting, but the gold standard used for grade prediction was not discussed; this may lead us to the conclusion that biopsy was part of the gold standard. We have highlighted the controversies surrounding biopsies and, accordingly, the research may have been affected by such issues. There are some studies which have applied deep learning for the prediction of tumour grade [75–77]. The AUC in these studies ranged from 77.0% to 88.2%. These results are not only worse than those obtained in the current research, but the Fuhrman grading system was also used as the gold standard.

4.2. Biopsy Grading and its Comparison with ML

A biopsy Biopsy is a commonly used diagnostic tool for the identification of RCC subtypes. The diagnostic accuracy of biopsy for RCC has been reported to range from 86.0 to 98.0%, but this can be influenced by various factors [13,78,79]. Notably, when it comes to grading RCC, the range of accuracy widens to between 43.0 and 76.0% [13,78–85]. Nevertheless, a biopsy's accuracy in classifying renal cell tumours is debatable (Millet et al., 2012). Different studies have contended that a kidney biopsy typically understates the final grade. For instance, biopsies underestimated the nuclear grade in 55% of instances and only properly identified 43% of the final nuclear grades [80]. In particular, the final nuclear grade was marginally more likely to be understated in biopsies of larger tumours, while histologic subtype analysis yielded more accurate results; especially when evaluating clear cell renal tumours. In the research by Blumenfeld et al. [80], only one case of the nuclear grade being over-estimated was reported. In the study of Millet et al. [82], biopsy led to under-estimation of the grade in 13 cases while, in 2 cases, it over-estimated the grade.

In our study, we found that the accuracy of biopsy was 35.7% in determining the tumour grade, with a sensitivity and specificity of 9.1% and 52.9%, respectively, in the 28 NHS samples (cohort 4) when nephrectomy was used as a gold standard as shown in Table 3. These results are in agreement with previous studies [80,82] which determined biopsy to perform poorly in grading tumours. The results obtained through biopsies were compared to our ML models, and the models out-performed biopsies by far; in fact, our worst-performing model was still better than biopsy. The best model had an accuracy of 96.4%, sensitivity of 90.9%, and specificity of 100% in the internal validation, comprising a 60.7% improvement in accuracy. Likewise, in the external validation, there was a 46.4% improvement in accuracy, with an accuracy, sensitivity, and specificity of 82.1%, 72.7%, and 88.2%, respectively as presented in Table A11. Therefore, we can conclude that ML approaches are able to distinguish low- from high-grade ccRCC with better accuracy, when compared to biopsy, and thus should be considered as a replacement.

In previous research, no paper has tackled the effect of tumour sub-region with regard to the grading of ccRCC; hence, there were no studies with which our results could be compared. The current research dived deeper into the possibility of pre-operatively grading ccRCC without the need for biopsy. Moreover, the effect of the information contained in different tumour sub-regions on grading was analysed. It is the belief of the authors of this research that the results of this study will assist clinicians in finding the best management strategies for patients of ccRCC, as well as enabling informative pre-treatment assessments that allow treatments to be tailored to individual patients.

5. Summary and Conclusions

5.0.1. *Limitations and Future research*

The work encountered a few challenges which are important to highlight. The samples used in this study were obtained from different institutions, and the scans were captured using different

 doi:10.20944/preprints202312.1379.v5

scanners and protocols. This may have lowered the overall performance of the models. However, it was important to use such data as the research was not meant to be institution-specific but, instead, generally applicable. Second, the retrospective nature of the research may have limited our work, and it is therefore recommended that more research should be conducted through a prospective study. Third, the current research assumed that the divided tumour sub-regions (25%, 50%, and 75% core and periphery) are heterogeneous in nature. In this regard, more research using pixel intensity measurement from different tumour sub-regions is encouraged. Fourth, manual segmentation is not only time-consuming but also subject to observer variability; thus, research on automated tumour image segmentation techniques is encouraged. Fifth, the predominant approach to grading ccRCC studies revolves around utilising a binarised model output. This is motivated by two primary factors. First, there exists a notable discrepancy in the sample sizes across different grades, with grades 1 and 4 exhibiting smaller sample size compared to grades 2 and 3. Second, adopting a 4-class model is perceived to yield minimal impact on patient management, given the similarity in management strategies between low grades (I and II) and high grades (III and IV) [86]. Nonetheless, there is merit in exploring the application of a 4-class model in forthcoming investigations, as doing so may validate the suitability of radiomics machine learning analyses in delineating distinct WHO/ISUP grading categories. Moreover, despite this being one of the few studies which has used a large sample size, we still consider our sample size to be small with respect to ML and AI approaches, which often require larger data sets for training. Finally, it's advisable to undertake a deep learning research using a substantial data set based on the WHO/ISUP grading systems.

5.1. Take-Home messages:

- *Radiomics features combined with ML algorithms have the potential to predict the WHO/ISUP grade of ccRCC more accurately than pre-operative biopsy.*
- *Analysing different tumor subregions, such as the 50% tumor core and 25% tumor periphery, provides valuable information for determining tumor grade.*
- *Analysing different cohorts from both single and multi-centre studies represented the effect of data heterogeneity on the model's performance. This underscores the importance of implementing a robust model that generalizes well for real-world applications in grading ccRCCs.*
- *The study highlighted the promising application of advanced imaging techniques and ML in oncology for precise tumor grading.*

5.2. Summary and Conclusions

In this study, an in-depth radiomics ML analysis of ccRCC was carried out with the purpose of determining the clinical significance of intra-tumoural sub-region heterogeneity in CT scans and biopsies with respect to the accuracy of tumour grading. In this regard, the results support the assertion that tumour sub-regions are an important factor to consider while grading ccRCC. We were able to demonstrate that the 50% tumour core can be considered the best sub-region for determining the tumour grade; however, this should not be interpreted as indicating that other tumour sub-regions are unimportant. Indeed, the results indicated only small differences in performance for the different tumour sub-regions; therefore, the different regions should be analysed independently and taken into consideration for the final grading outcome. Regarding the second objective on the importance of biopsy in grading, through comparison of our research results with biopsy results, we were able to demonstrate that ML approaches yield much better results in terms of determining the ccRCC WHO/ISUP grade. Finally, the performance of the ML models in determining tumour grade demonstrates the potential benefit of using ML as an alternative or replacement for biopsy in determining the tumour grade.

In conclusion, the present work demonstrated the potential of ML models for distinguishing low- from high-grade ccRCC. In essence, ML approaches can act as a "virtual biopsy," being potentially far superior to biopsy for grading purposes. These findings have important clinical significance for

 doi:10.20944/preprints202312.1379.v5

addressing the challenges that are experienced in relation to biopsies, leading to improved clinical management and contributing to oncological precision medicine.

Author Contributions: Conceptualisation, Abeer J. Alhussaini, Ghulam Nabi, and J. Douglas Steele; Data curation, Abeer J. Alhussaini and Ghulam Nabi; Formal analysis, Abeer J. Alhussaini; Investigation, Abeer J. Alhussaini; Methodology, Abeer J. Alhussaini; Project administration, Abeer J. Alhussaini, Ghulam Nabi, and J. Douglas Steele; Resources, Abeer J. Alhussaini, Ghulam Nabi, and J. Douglas Steele; Software, Abeer J. Alhussaini; Second observer segmentation, Adel Jawli; Supervision, Ghulam Nabi and J. Douglas Steele; Validation, Abeer J. Alhussaini; Visualisation, Abeer J. Alhussaini; Writing—original draft, Abeer J. Alhussaini; Writing—review and editing, Abeer J. Alhussaini, Ghulam Nabi, and J. Douglas Steele. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: In the cohort 1 retrospective study, approval for the research was obtained, including the experiments, access to clinical follow-up data, and study protocols, from the NHS Tayside Ninewells Hospital Medicine School in Dundee, Scotland, under the East of Scotland Ethical Committee and Caldicott approval number (IGTCAL11334), dated 21 October 2022. This study adhered to the Declaration of Helsinki. Informed consent for the research was not required, as CT scan image acquisition is a routine examination procedure for patients suspected of having ccRCC.

Data Availability Statement: The data provided in the cohort 1 study are available on request from the corresponding authors. For cohort 2, the data are readily available from the Kits GitHub page and Cancer Imaging Archive (CIA) [30–32]. The codes used to reproduce the results can be found on GitHub, upon request, at this link.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Representation of Cohort 1 diagnostic performance using core sub-regions and 100% tumour under different models. In the training set, the sample size was 125, with 76 high-grade and 49 low-grade samples. In the testing/internal validation set, the sample size was 62, comprising 31 high-grade and 31 low-grade samples.

COHORT 1 CORE N=187 INTERNAL VALIDATION														
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	71.0±11	74.2±11	71.0±11	72.6±11	69.4±11	74.2±11	74.2±11	71.0±11	74.2±11	80.7±10	69.4±11	72.9	
	SPE	71.0±16	71.0±16	71.0±16	67.7±16	71.0±16	80.7±14	71.0±16	67.7±16	71.0±16	80.7±14	74.2±15	72.4	
	SEN	71.0±16	77.4±15	71.0±16	77.4±15	67.7±16	67.7±16	77.4±15	74.2±15	77.4±15	80.7±14	64.5±17	73.3	
	AUC	71.0±11	74.0±11	71.0±11	73.0±11	69.0±12	74.0±11	74.0±11	71.0±11	74.0±11	**81.0±10**	69.0±12	72.8	7
	MCC	0.42	0.48	0.42	0.45	0.39	0.49	0.48	0.42	0.48	0.61	0.39	-	
	F1	0.71	0.75	0.71	0.74	0.69	0.72	0.75	0.72	0.75	0.81	0.68	-	
	MeN	1.00	0.62	1.00	0.47	0.82	0.32	0.62	0.64	0.62	0.10	0.49	-	
	χ^2	0.13	0.05	0.13	0.07	0.18	0.04	0.05	0.12	0.05	0	0.16	-	
50%	ACC	79.0±10	79.0±10	82.3±10	74.2±11	74.2±11	75.8±11	77.4±10	79.0±10	79.0±10	82.3±10	75.8±11	78.0	
	SPE	74.2±15	74.2±15	87.1±12	71.0±16	74.2±15	77.4±15	74.2±15	74.2±15	71.0±16	77.4±15	71.0±16	73.9	
	SEN	83.9±13	83.2±13	77.4±15	77.4±15	74.2±15	74.2±15	80.7±14	83.9±13	87.1±12	87.1±12	80.7±14	79.6	
	AUC	79.0±10	79.0±10	82.0±10	74.0±11	74.0±11	76.0±11	77.0±11	79.0±10	79.0±10	**82.0±10**	76.0±11	**77.9**	8
	MCC	0.58	0.58	0.65	0.48	0.48	0.52	0.55	0.58	0.59	0.65	0.52	-	
	F1	0.80	0.80	0.81	0.75	0.74	0.75	0.78	0.80	0.81	0.8	0.77	-	
	MeN	0.41	0.41	0.37	0.62	1.00	0.80	0.59	0.41	0.17	0.37	0.44	-	
	χ^2	0.01	0.01	0	0.05	0.05	0.03	0.02	0.01	0.01	0	0.03	-	
25%	ACC	74.2±11	74.2±11	79.0±10	74.2±11	72.3±11	74.2±11	77.4±10	79.0±10	80.7±10	79.0±10	72.6±11	76.1	
	SPE	71.0±16	74.2±15	71.0±16	74.2±15	67.7±16	71.0±16	74.2±15	71.0±16	77.4±15	77.4±15	67.7±16	72.4	
	SEN	77.4±15	74.2±15	87.1±12	74.2±15	77.4±15	77.4±15	80.7±14	87.1±12	83.9±13	80.7±14	77.4±15	78.1	
	AUC	74.0±11	74.0±11	79.0±10	74.0±11	73.0±11	74.0±11	77.0±11	79.0±10	**81.0±10**	79.0±10	73.0±11	76.1	19
	MCC	0.48	0.48	0.59	0.48	0.45	0.48	0.55	0.59	0.61	0.58	0.45	-	
	F1	0.75	0.74	0.81	0.74	0.74	0.75	0.78	0.81	0.81	0.79	0.74	-	
	MeN	0.62	1.00	0.17	1.00	0.47	0.62	0.59	0.17	0.56	0.78	0.47	-	
	χ^2	0.05	0.05	0.01	0.05	0.07	0.05	0.02	0.01	0.00	0.01	0.07	-	
AVG	AUC	74.7	75.7	77.3	73.7	72.0	74.7	76.0	76.3	78.0	**80.7**	72.7	-	
100%	ACC	77.4±10	79.0±10	79.0±10	74.2±11	69.4±11	72.0±11	75.8±11	75.8±11	74.2±11	85.5±9	77.4±10	76.3	
	SPE	77.4±15	71.0±16	77.4±15	71.0±16	71.0±16	71.0±16	74.2±15	74.2±15	77.4±15	77.4±15	74.2±15	74.2	
	SEN	77.4±15	87.1±12	80.7±14	77.4±15	67.7±16	74.2±15	77.4±15	77.4±15	71.0±16	93.6±8	80.7±14	78.6	
	AUC	77.0±11	79.0±11	79.0±10	74.0±11	69.0±12	73.0±11	76.0±11	76.0±11	74.0±11	**85.0±9**	77.0±11	**76.3**	16
	MCC	0.55	0.59	0.58	0.48	0.39	0.45	0.52	0.52	0.48	0.72	0.55	-	
	F1	0.77	0.81	0.79	0.75	0.69	0.73	0.76	0.76	0.73	0.87	0.78	-	
	MeN	1.00	0.17	0.78	0.62	0.82	0.81	0.80	0.80	0.62	0.10	0.59	-	
	χ^2	0.02	0.01	0.01	0.05	0.18	0.08	0.03	0.03	0.05	0	0.02	-	

 doi:10.20944/preprints202312.1379.v5

Table A2. Representation of cohort 1 diagnostic performance using periphery sub-regions under different models. In the training set, the sample size was 125, with 76 high-grade and 49 low-grade samples. In the testing/internal validation set, the sample size was 62, comprising 31 high-grade and 31 low-grade samples.

COHORT 1 PERIPHERY N=187 INTERNAL VALIDATION														
REGION		**SVM**	**RF**	**XGB**	**NB**	**MLP**	**LSTM**	**LR**	**QDA**	**LGB**	**CB**	**ADB**	**AVG**	**FS**
	ACC	74.2±11	74.2±11	71.0±11	75.8±11	71.0±11	69.0±12	71.0±11	74.2±11	71.0±11	77.4±10	72.6±11	72.8	
	SPE	71.0±16	71.0±16	71.0±16	80.7±14	71.0±16	77.4±15	67.7±16	67.7±16	71.0±16	77.4±15	71.0±16	72.4	
	SEN	77.4±15	77.4±15	71.0±16	71.0±16	71.0±16	61.3±17	74.2±15	80.7±14	71.0±16	77.4±15	74.2±15	73.3	
75%	**AUC**	74.0±11	74.0±11	71.0±11	76.0±11	71.0±11	69.0±12	71.0±11	74.0±11	71.0±11	**77.0±11**	73.0±11	72.8	8
	MCC	0.48	0.48	0.42	0.52	0.42	0.39	0.42	0.49	0.42	0.55	0.45	-	
	F1	0.75	0.75	0.71	0.75	0.71	0.67	0.72	0.76	0.71	0.77	0.73	-	
	McN	0.62	0.62	1.00	0.44	1.00	0.25	0.64	0.32	1.00	1.00	0.81	-	
	χ^2	0.05	0.05	0.13	0.03	0.13	0.12	0.12	0.04	0.13	0.02	0.08	-	
	ACC	71.0±11	75.8±11	79.0±10	72.6±11	72.6±11	69.4±11	74.2±11	74.2±11	74.2±11	77.4±10	71.0±11	73.8	
	SPE	71.0±16	74.2±15	71.0±16	71.0±16	71.0±16	64.5±17	74.2±15	64.5±17	61.3±17	77.4±15	64.5±17	69.5	
	SEN	71.0±16	77.4±15	87.1±12	74.2±15	74.2±15	74.2±15	74.2±15	83.9±13	87.1±12	77.4±15	77.4±15	76.0	
50%	**AUC**	71.0±11	76.0±11	**79.0±10**	73.0±11	73.0±11	69.0±12	74.0±11	74.0±11	74.0±11	77.0±11	71.0±11	73.7	11
	MCC	0.42	0.52	0.59	0.45	0.45	0.39	0.48	0.49	0.50	0.55	0.42	-	
	F1	0.71	0.76	0.81	0.73	0.73	0.71	0.74	0.76	0.77	0.77	0.72	-	
	McN	1.00	0.80	0.17	0.81	0.81	0.49	1.00	0.13	0.05	1.00	0.35	-	
	χ^2	0.13	0.03	0.01	0.08	0.08	0.16	0.05	0.02	0.01	0.02	0.09	-	
	ACC	77.4±10	80.7±10	79.0±10	82.3±10	75.8±10	72.6±11	77.4±10	80.7±10	79.0±10	83.9±9	75.8±11	78.6	
	SPE	74.2±15	80.7±14	71.0±16	83.9±13	80.7±13	71.0±16	77.4±15	74.2±15	74.2±15	80.7±14	64.5±17	75.7	
	SEN	80.7±14	89.7±11	87.1±12	80.7±14	77.4±15	74.2±15	77.4±15	87.1±12	83.9±13	87.1±12	87.1±12	81.4	
25%	**AUC**	77.0±11	81.0±10	79.0±10	82.0±10	76.0±11	73.0±11	77.0±11	81.0±10	79.0±10	**84.0±9**	76.0±11	**78.6**	9
	MCC	0.55	0.61	0.59	0.65	0.52	0.45	0.55	0.62	0.58	0.68	0.53	-	
	F1	0.78	0.81	0.81	0.82	0.77	0.73	0.77	0.82	0.80	0.84	0.78	-	
	McN	0.59	1.00	0.17	0.76	0.44	0.81	1.00	0.25	0.41	0.53	0.07	-	
	χ^2	0.02	0	0.01	0	0.03	0.08	0.02	0.00	0.01	0	0.01	-	
AVG	**AUC**	74.0	77.0	76.3	77.0	73.3	70.3	74.0	76.3	78.0	**79.3**	73.3	-	

Table A3. Representation of cohort 2 diagnostic performance using core sub-regions and 100% tumour under different models. The sample size in the training set was 136 (high-grade=56, low-grade=80), and in the testing/internal validation set, it was 68 (high-grade=21, low-grade=47).

COHORT 2 CORE N=204 INTERNAL VALIDATION														
REGION		**SVM**	**RF**	**XGB**	**NB**	**MLP**	**LSTM**	**LR**	**QDA**	**LGB**	**CB**	**ADB**	**AVG**	**FS**
	ACC	86.8±8	85.3±8	82.3±9	80.9±9	82.0±9	82.3±9	85.3±8.4	85.3±8	80.9±9	86.8±8	77.9±10	83.5	
	SPE	93.6±7	85.1±10	78.7±12	80.9±11	83.0±11	83.0±11	89.4±9	89.4±9	76.6±12	87.2±10	76.6±12	84.9	
	SEN	71.4±19	85.7±15	90.5±11	81.0±17	81.0±17	81.0±17	76.2±18	76.2±18	90.5±11	85.7±15	81.0±17	81.8	
75%	**AUC**	83.0±9	85.0±9	85.0±9	81.0±9	82.0±9	82.0±9	83.0±8.9	83.0±9	84.0±9	**86.0±8**	79.0±10	83.0	9
	MCC	0.68	0.68	0.65	0.59	0.61	0.61	0.66	0.66	0.62	0.71	0.54	-	
	F1	0.77	0.78	0.76	0.72	0.74	0.74	0.76	0.76	0.75	0.80	0.69	-	
	McN	0.32	0.21	0.02	0.17	0.25	0.25	1.00	1.00	0.01	0.32	0.07	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
	ACC	88.2±8	88.2±8	88.2±8	88.2±8	85.0±8	83.8±9	88.2±8	91.2±7	88.2±8	89.7±7	86.8±8	87.9	
	SPE	91.5±8	87.2±10	87.2±10	91.5±8	85.1±10	80.9±11	95.8±5	91.5±8	85.1±10	89.4±8.8	85.1±10	88.2	
	SEN	81.0±17	90.5±11	90.5±11	90.0±11	85.7±15	90.5±11	71.4±19	90.5±11	95.2±7	90.5±11	90.5±11	87.8	
50%	**AUC**	86.0±8.2	89.0±7.4	89.0±7.4	86.0±8.2	85.0±8.5	86.0±8.2	84.0±8.7	91.0±6.8	90.0±7.1	**90.0±7.1**	88.0±7.7	**87.6**	12
	MCC	0.72	0.74	0.74	0.72	0.68	0.67	0.72	0.80	0.76	0.77	0.72	-	
	F1	0.81	0.83	0.83	0.81	0.78	0.78	0.79	0.86	0.83	0.84	0.81	-	
	McN	1.00	0.16	0.16	1.00	0.21	0.03	0.16	0.41	0.03	0.26	0.1	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
	ACC	85.3±8	85.3±8	82.4±9	82.4±9	79.0±10	79.4±10	82.4±9	83.8±9	80.0±10	85.3±8	85.3±8	83.5	
	SPE	85.1±10	85.1±10	80.9±11	83.0±11	78.7±12	80.9±11	83.0±11	83.0±11	83.0±11	83.0±11	87.2±10	83.0	
	SEN	85.7±15	85.7±15	85.7±15	81.0±17	81.0±17	76.2±18	81.0±17	85.7±15	81.0±17	85.7±15	81.0±17	82.7	
25%	**AUC**	85.0±9	85.0±9	83.0±9	82.0±9	80.0±10	79.0±10	82.0±9	84.0±9	81.0±9	**85.0±9**	84.0±9	82.7	10
	MCC	0.68	0.68	0.63	0.61	0.56	0.55	0.61	0.65	0.59	0.68	0.67	-	
	F1	0.78	0.78	0.75	0.74	0.71	0.70	0.74	0.77	0.72	0.78	0.77	-	
	McN	0.21	0.21	0.08	0.25	0.11	0.29	0.25	0.13	0.17	0.21	0.53	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
AVG	**AUC**	84.7	86.3	85.7	83.0	82.3	82.3	83.0	86.0	85.0	87.0	83.7	-	
	ACC	85.3±8	85.3±8	80.9±9	85.3±8	82.0±9	82.4±9	82.4±9	82.4±9	86.8±8	86.8±8	83.8±9	84.1	
	SPE	91.5±8	85.1±10	83.0±11	87.2±10	83.0±11	85.1±10	81.0±11	81.0±11	89.4±9	87.2±10	85.1±10	85.3	
	SEN	71.4±19	85.7±15	76.2±18	81.0±17	81.0±17	76.2±18	85.7±14	85.7±14	81.0±17	85.7±15	81.0±17	81.0	
100%	**AUC**	81.0±9	85.0±9	80.0±10	84.0±9	82.0±9	81.0±9	83.0±9	83.0±9	85.0±9	**86.0±8**	83.0±9	**83.0**	9
	MCC	0.65	0.68	0.57	0.67	0.61	0.60	0.63	0.63	0.69	0.71	0.64	-	
	F1	0.75	0.78	0.71	0.77	0.74	0.73	0.75	0.75	0.79	0.80	0.76	-	
	McN	0.53	0.21	0.41	0.53	0.25	0.56	0.08	0.08	0.74	0.32	0.37	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	

 doi:10.20944/preprints202312.1379.v5

Table A4. Representation of cohort 2 diagnostic performance using periphery sub-regions under different models. The sample size in the training set was 136 (high-grade=56, low-grade=80), and in the testing/internal validation set, it was 68 (high-grade=21, low-grade=47).

COHORT 2 PERIPHERY N=204 INTERNAL VALIDATION													
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	88.2±8	85.3±8	83.8±9	86.8±8	79.4±10	76.5±10	76.5±10	86.8±8	80.9±9	85.3±8	88.2±8	83.8	
	SPE	95.8±5	82.4±11	85.1±10	91.5±8	76.6±12	72.3±13	89.4±9	93.6±7	80.9±11	87.2±10	87.2±10	86.8	
	SEN	71.4±19	85.7±15	81±17	76.2±18	85.7±15	85.7±15	71.4±19	71.4±19	81.0±17	81.0±17	76.2±18	78.8	
	AUC	84.0±9	**85.0±9**	83.0±9	84.0±9	81.0±9	79.0±10	80.0±10	83.0±9	81.0±9	84.0±9	82.0±9.1	**82.4**	8
	MCC	0.72	0.68	0.64	0.69	0.58	0.54	0.62	0.68	0.59	0.67	0.63	-	
	F1	0.79	0.78	0.76	0.78	0.72	0.69	0.73	0.77	0.72	0.77	0.74	-	
	McN	0.16	0.21	0.37	0.74	0.03	0.01	0.76	0.32	0.17	0.53	0.76	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
50%	ACC	86.8±8	88.2±8	75.0±10	85.3±8	75.0±10	79.4±10	76.5±10	88.2±8	76.5±10	91.2±7	88.2±8	85.1	
	SPE	91.5±8	95.8±5	70.2±13	87.2±10	70.2±13	78.7±12	76.6±12	95.8±5	72.3±13	91.5±8	95.8±5	84.9	
	SEN	76.2±18	71.4±19	85.7±15	81.0±17	85.7±15	81.0±17	76.2±18	71.4±19	85.7±15	90.5±11	71.4±19	79.7	
	AUC	84.0±9	84.0±9	78.0±10	84.0±9	78.0±10	80.0±10	76.0±10	84.0±9	79.0±10	**91.0±7**	84.0±9	82	12
	MCC	0.69	0.72	0.52	0.67	0.52	0.56	0.50	0.72	0.54	0.80	0.72	-	
	F1	0.78	0.79	0.68	0.77	0.68	0.71	0.67	0.79	0.69	0.86	0.79	-	
	McN	0.74	0.16	0.01	0.53	0.01	0.11	0.13	0.16	0.01	0.41	0.16	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
25%	ACC	85.3±8	85.3±8	76.5±10	82.4±9	77.9±10	75.0±10	76.5±10	89.7±7	80.9±9	83.8±9	85.3±8	82.4	
	SPE	85.1±10	87.2±10	72.3±13	83.0±11	76.6±12	74.5±12	78.7±12	93.6±7	83.0±11	85.1±10	85.1±10	82.8	
	SEN	85.7±15	81.0±17	85.7±15	81.0±17	81.0±17	76.2±18	71.4±19	81.0±17	76.2±18	81.0±17	85.7±15	80.52	
	AUC	85.0±9	84.0±9	79.0±10	82.0±9	79.0±10	75.0±10	75.0±10	**87.0±8**	80.0±10	83.0±9	85.0±9	81.27	16
	MCC	0.68	0.67	0.54	0.61	0.54	0.48	0.48	0.76	0.57	0.64	0.68	-	
	F1	0.78	0.77	0.69	0.74	0.69	0.65	0.65	0.83	0.71	0.76	0.78	-	
	McN	0.21	0.53	0.01	0.25	0.07	0.09	0.32	0.71	0.41	0.37	0.21	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
AVG	AUC	84.3	84.3	80.0	83.3	79.3	78.0	77.0	84.7	80.0	**86.0**	83.7	-	

Table A5. Representation of cohort 3 diagnostic performance using core sub-regions and 100% tumour under different models. The sample size in the training set was 261 (high-grade=122, low-grade=139), and in the testing/internal validation set, it was 130 (high-grade=62, low-grade=68).

COHORT 3 CORE N=391 INTERNAL VALIDATION													
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	70.8±8	76.9±7	74.6±7	71.5±8	73.1±8	71.5±8	73.1±8	74.6±7	73.1±8	76.2±7	73.9±8	73.6	
	SPE	69.1±11	76.5±10	73.5±10	73.5±10	77.9±10	72.1±11	70.6±11	77.9±10	76.5±10	77.9±10	76.8±10	74.7	
	SEN	72.6±11	77.4±10	75.8±11	69.4±11	67.7±12	71.0±11	75.8±11	71.0±11	69.4±11	74.2±11	71.0±11	72.3	
	AUC	71.0±8	**77.0±7**	75.0±7	71.0±8	73.0±8	72.0±8	73.0±8	74.0±8	73.0±8	76.0±7	74.0±8	73.6	14
	MCC	0.42	0.54	0.49	0.43	0.46	0.43	0.46	0.49	0.46	0.52	0.48	-	
	F1	0.70	0.76	0.74	0.70	0.71	0.70	0.73	0.73	0.71	0.75	0.72	-	
	McN	0.52	0.72	0.60	0.87	0.40	0.87	0.40	0.60	0.61	0.86	0.73	-	
	χ^2	0.01	0	0	0	0	0.01	0	0	0	0	0	-	
50%	ACC	76.9±7	76.9±7	76.2±7	76.9±7	73.1±8	76.2±7	77.7±7	79.2±7	78.5±7	80.0±7	74.6±7	76.6	
	SPE	75.0±10	77.9±10	77.9±10	77.9±10	70.6±11	72.1±11	76.5±10	76.5±10	77.9±10	79.4±10	76.5±10	76.3	
	SEN	79.0±10	75.8±11	74.2±11	75.8±11	75.8±11	80.7±10	79.0±10	82.3±10	79.0±10	80.7±10	72.6±11	77.7	
	AUC	77.0±7	77.0±7	76.0±7	77.0±7	73.0±8	76.0±7	78.0±7	79.0±7	78.0±7	**80.0±7**	75.0±7	**76.9**	24
	MCC	0.54	0.54	0.52	0.54	0.46	0.53	0.55	0.59	0.57	0.60	0.49	-	
	F1	0.77	0.76	0.75	0.76	0.73	0.76	0.77	0.79	0.78	0.79	0.73	-	
	McN	0.47	1.00	0.86	1.00	0.40	0.21	0.58	0.34	0.71	0.69	0.86	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
25%	ACC	74.6±7	74.6±7	72.1±8	73.1±8	72.3±8	71.5±8	71.5±8	73.1±8	73.1±8	76.2±7	70.8±8	73.0	
	SPE	80.9±9	75.0±10	75.0±10	72.1±11	76.5±10	72.1±11	71.0±11	72.2±10	73.5±10	77.9±10	72.1±11	74.2	
	SEN	67.7±11	74.2±11	69.4±11	74.2±11	67.7±12	71.0±11	72.6±11	74.2±11	72.6±11	74.2±11	69.4±11	71.6	
	AUC	74.0±8	75.0±7	72.0±8	73.0±8	72.0±8	72.0±8	72.0±8	73.0±8	73.0±8	**76.0±7**	71.0±8	73.0	32
	MCC	0.49	0.49	0.44	0.46	0.44	0.43	0.43	0.46	0.46	0.52	0.41	0.46	
	F1	0.72	0.74	0.70	0.73	0.70	0.70	0.71	0.72	0.72	0.75	0.69	-	
	McN	0.22	0.86	0.74	0.61	0.50	0.87	0.62	0.61	0.87	0.86	1.00	-	
	χ^2	0	0	0	0	0	0.01	0.01	0	0	0	0.01	-	
AVG	AUC	74.0	76.3	74.3	73.7	72.7	73.3	74.3	75.3	74.7	**77.3**	73.3	-	
100%	ACC	73.9±8	76.2±7	75.4±7	75.4±7	71.5±8	73.9±8	74.6±7	75.4±7	78.5±7	78.5±7	74.6±7	75.2	
	SPE	73.5±10	72.1±11	77.9±10	79.4±10	76.5±10	72.1±11	71.0±11	76.5±10	82.4±9	80.9±9	76.5±10	76.2	
	SEN	74.2±11	80.7±10	72.6±11	71.0±11	66.1±12	75.8±11	79.0±10	74.2±11	74.2±11	75.8±11	72.6±11	74.2	
	AUC	74.0±8	76.0±7	75.0±7	75.0±7	71.0±8	74.0±8	75.0±7	75.0±7	78.0±7	**78.0±7**	75.0±7	**75.1**	15
	MCC	0.48	0.53	0.51	0.51	0.43	0.48	0.50	0.51	0.57	0.57	0.49	-	
	F1	0.73	0.76	0.74	0.73	0.69	0.73	0.75	0.74	0.77	0.77	0.73	-	
	McN	0.73	0.21	0.72	0.48	0.41	0.49	0.22	1.00	0.45	0.71	0.86	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	

 doi:10.20944/preprints202312.1379.v5

Table A6. Representation of cohort 3 diagnostic performance using periphery sub-regions under different models. The sample size in the training set was 261 (High-grade=122, Low-grade=139), and in the testing/internal validation set, it was 130 (High-grade=62, Low-grade=68).

COHORT 3 PERIPHERY N=391 INTERNAL VALIDATION													
REGION		**SVM**	**RF**	**XGB**	**NB**	**MLP**	**LSTM**	**LR**	**QDA**	**LGB**	**CB**	**ADB**	**AVG**	**FS**
	ACC	70.0±8	80.0±7	74.6±7	71.5±8	70.0±8	70.8±8	72.3±8	75.4±8	73.1±7	77.7±7	72.3±8	73.5	
	SPE	70.6±11	82.4±9	73.5±10	72.1±11	70.6±11	69.1±11	70.6±11	77.9±10	76.5±10	77.9±10	72.1±11	73.9	
	SEN	69.4±11	77.4±10	75.8±11	71.0±11	69.4±11	72.6±11	74.2±11	72.6±11	69.4±11	77.4±10	72.6±11	72.9	
75%	**AUC**	70.0±8	**80.0±7**	75.0±7	72.0±8	70.0±8	71.0±8	72.0±8	75.0±7	73.0±8	78.0±7	72.0±8	73.5	22
	MCC	0.40	0.60	0.49	0.43	0.40	0.42	0.45	0.51	0.46	0.55	0.45	-	
	F1	0.69	0.79	0.74	0.70	0.69	0.70	0.72	0.74	0.71	0.77	0.71	-	
	McN	0.87	0.69	0.60	0.87	0.87	0.52	0.50	0.72	0.61	0.85	0.74	-	
	χ^2	0.01	0	0	0.01	0.01	0.01	0	0	0	0	0	-	
	ACC	73.9±8	76.9±7	75.4±7	72.3±8	70.0±8	71.5±8	75.4±7	74.6±7	74.6±7	76.2±7	69.2±8	74.0	
	SPE	70.6±11	77.9±10	75.0±10	72.1±11	75.0±10	69.1±11	69.1±11	80.9±10	69.1±11	70.6±11	66.2±11	71.7	
	SEN	77.4±10	75.8±11	75.8±11	72.6±11	64.5±12	74.2±11	82.3±10	67.7±12	80.7±10	82.3±10	72.6±11	75.1	
50%	**AUC**	74.0±8	**77.0±7**	75.0±7	72.0±8	70.0±8	72.0±8	76.0±7	74.0±8	75.0±7	76.0±7	69.0±8	73.6	26
	McC	0.48	0.54	0.51	0.45	0.40	0.43	0.52	0.49	0.50	0.53	0.39	-	
	F1	0.74	0.76	0.75	0.71	0.67	0.71	0.76	0.72	0.75	0.77	0.69	-	
	McN	0.30	1.00	0.72	0.74	0.42	0.41	0.08	0.22	0.12	0.11	0.34	-	
	χ^2	0	0	0	0	0.01	0	0	0	0	0	0.01	-	
	ACC	75.4±8	77.7±7	76.9±7	75.4±7	73.1±8	73.1±8	75.4±7	75.4±7	76.2±7	76.2±7	74.6±7	75.4	
	SPE	83.8±11	77.9±10	80.9±9	76.5±10	72.1±11	73.5±10	76.5±10	72.1±11	77.9±10	73.5±10	75.0±10	76.5	
	SEN	66.1±12	77.4±10	72.6±11	74.2±11	74.2±11	72.6±11	74.2±11	79.0±10	74.2±11	79.0±10	74.2±11	74.3	
25%	**AUC**	75.0±7	**78.0±7**	77.0±7	75.0±7	73.0±8	73.0±8	75.0±7	76.0±7	76.0±7	76.0±7	75.0±7	**75.4**	21
	MCC	0.51	0.55	0.54	0.51	0.46	0.46	0.51	0.51	0.52	0.53	0.49	-	
	F1	0.72	0.77	0.75	0.74	0.72	0.72	0.74	0.75	0.75	0.76	0.74	-	
	McN	0.08	0.85	0.47	1.00	0.61	0.87	1.00	0.29	0.86	0.37	0.86	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
AVG	**AUC**	73.0	**78.3**	75.7	73.0	71.0	72.0	74.3	75.0	74.7	76.7	72.0	-	

Table A7. Representation of diagnostic performance for the external validation of cohort 1 using core sub-regions under different models. Cohort 2, comprising 204 samples (77 high-grade and 127 low-grade), was used as the training set, while cohort 1, with 187 samples (107 high-grade and 80 low-grade), served as the testing/external validation set.

Cohort 1 CORE N=187 EXTERNAL VALIDATION													
REGION		**SVM**	**RF**	**XGB**	**NB**	**MLP**	**LSTM**	**LR**	**QDA**	**LGB**	**CB**	**ADB**	**AVG**	**FS**
	ACC	66.3±7	65.0±7	65.0±7	62.0±7	61.0±7	63.1±7	64.7±7	66.3±7	62.0±7	65.7±7	62.5±7	63.9	
	SPE	61.3±11	47.5±11	55.0±11	55.0±7	53.8±11	53.7±11	63.8±11	68.8±10	47.5±11	58.8±11	45.0±11	56.3	
	SEN	70.1±9	77.6±8	72.0±9	67.2±9	66.3±9	70.1±9	65.4±9	64.5±9	73.0±8	71.0±9	75.7±8	70.3	
75%	**AUC**	66.0±7	63.0±7	63.0±7	61.0±7	60.0±7	62.0±7	65.0±7	**67.0±7**	60.0±7	65.0±7	60.0±7	62.9	19
	MCC	0.31	0.26	0.27	0.22	0.20	0.24	0.29	0.33	0.21	0.30	0.22	-	
	F1	0.7	0.72	0.70	0.67	0.66	0.68	0.68	0.69	0.69	0.70	0.70	-	
	McN	0.90	0.03	0.46	0.91	0.91	0.55	0.32	0.10	0.12	0.80	0.03	-	
	χ^2	0.01	0	0.01	0.04	0.05	0.01	0.02	0.01	0	0.01	0	-	
	ACC	68.5±7	64.7±7	62.6±7	63.1±7	61.0±7	62.6±7	68.5±7	70.1±7	64.7±7	68.5±7	63.1±7	65.2	
	SPE	57.5±11	57.5±11	58.8±11	63.8±11	51.3±11	63.8±11	62.5±11	62.5±11	55.0±11	60.0±11	47.5±11	58.2	
	SEN	76.6±8	70.1±9	65.4±9	62.6±9	68.2±9	61.7±9	72.9±8	75.7±8	72.0±9	74.8±8	74.7±8	70.4	
50%	**AUC**	67.0±7	64.0±7	62.0±7	63.0±7	60.0±7	63.0±7	68.0±7	**69.0±7**	63.0±7	67.0±7	61.0±7	**64.3**	20
	MCC	0.35	0.28	0.24	0.26	0.20	0.25	0.35	0.38	0.27	0.35	0.23	-	
	F1	0.74	0.69	0.67	0.66	0.67	0.65	0.73	0.74	0.70	0.73	0.70	-	
	McN	0.24	0.81	0.63	0.19	0.56	0.15	0.90	0.59	0.25	0.52	0.07	-	
	χ^2	0	0.01	0.04	0.04	0.03	0.05	0	0	0.01	0	0	-	
	ACC	66.8±7	64.7±7	68.5±7	64.7±7	62.0±7	64.7±7	67.4±7	65.8±7	63.1±7	65.8±7	63.1±7	65.1	
	SPE	63.8±11	53.8±11	55.0±11	66.3±10	50.0±11	57.5±11	66.3±10	63.8±11	52.5±11	60.0±11	45.0±11	57.6	
	SEN	69.2±9	72.9±8	78.5±8	63.6±9	71.0±9	70.1±9	68.2±9	67.3±9	71.0±9	70.1±9	76.6±8	70.6	
25%	**AUC**	66.0±7	63.0±7	67.0±7	65.0±7	61.0±7	64.0±7	**67.0±7**	66.0±7	62.0±7	65.0±7	61.0±7	**64.3**	15
	MCC	0.33	0.27	0.35	0.29	0.21	0.28	0.34	0.31	0.24	0.30	0.23	-	
	F1	0.70	0.70	0.74	0.67	0.68	0.69	0.71	0.69	0.69	0.70	0.70	-	
	McN	0.61	0.32	0.09	0.14	0.29	0.81	0.37	0.45	0.40	1.00	0.02	-	
	χ^2	0	0	0	0.02	0.01	0.01	0	0.01	0.01	0.01	0	-	
AVG	**AUC**	66.3	63.3	64.0	63.0	60.3	63.0	66.7	67.3	61.7	65.7	60.7	-	
	ACC	66.8±7	66.3±7	65.8±7	65.2±7	59.9±7	64.2±7	62.6±7	67.9±7	62.0±7	65.8±7	64.2±7	64.6	
	SPE	61.3±11	62.5±11	61.3±11	52.5±11	51.3±11	51.3±11	61.3±11	68.8±10	47.5±11	52.5±11	60.0±11	56.8	
	SEN	71.0±9	69.2±9	69.2±9	74.8±8	66.2±9	73.8±8	63.6±9	67.3±9	72.9±8	75.7±8	67.3±9	70.1	
100%	**AUC**	66.0±7	66.0±7	65.0±7	64.0±7	59.0±7	63.0±7	62.0±7	**68.0±7**	60.0±7	64.0±7	64.0±7	**63.7**	23
	MCC	0.32	0.32	0.30	0.28	0.18	0.26	0.25	0.36	0.21	0.29	0.27	-	
	F1	0.71	0.70	0.70	0.71	0.65	0.70	0.66	0.71	0.69	0.72	0.68	-	
	McN	1.00	0.71	0.80	0.17	0.73	0.18	0.34	0.20	0.12	0.13	0.71	-	
	χ^2	0	0.01	0.01	0	0.06	0	0.05	0	0	0	0.02	-	

 doi:10.20944/preprints202312.1379.v5

Table A8. Representation of diagnostic performance for the external validation of cohort 1 using periphery sub-regions under different models. The training set utilised Cohort 2, which consisted of 204 samples (77 high-grade and 127 low-grade), whereas Cohort 1, comprising 187 samples (107 high-grade and 80 low-grade), was employed for testing/external validation set.

Cohort 1 PERIPHERY N=187 EXTERNAL VALIDATION													
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	62.6±7	71.7±7	62.6±7	62.0±7	62.6±7	61.5±7	62.6±7	66.3±7	64.2±7	65.2±7	62.0±7	63.9	15
	SPE	57.5±11	53.8±11	48.8±11	51.3±11	47.5±11	53.8±11	61.3±11	53.8±11	50.0±11	51.3±11	48.8±11	52.5	
	SEN	66.4±9	71.0±9	72.90±8	70.1±9	73.8±8	67.3±9	63.6±9	75.7±8	74.8±8	75.7±8	72.0±9	71.2	
	AUC	62.0±7	62.0±7	61.0±7	61.0±7	61.0±7	61.0±7	62.0±7	**65.0±7**	62.0±7	63.0±7	60.0±7	61.8	
	MCC	0.24	0.25	0.22	0.22	0.22	0.21	0.25	0.30	0.26	0.28	0.21	-	
	F1	0.67	0.69	0.69	0.68	0.69	0.67	0.66	0.72	0.70	0.71	0.68	-	
	MeN	0.81	0.47	0.15	0.41	0.09	0.81	0.34	0.17	0.11	0.11	0.19	-	
	χ^2	0.04	0.01	0	0.02	0	0.04	0.05	0	0	0	0.01	-	
50%	ACC	61.5±7	66.3±7	64.7±7	64.2±7	61.5±7	63.1±7	62.6±7	67.4±7	63.1±7	65.2±7	65.2±7	64.1	21
	SPE	55.0±11	50.0±11	55.0±11	51.3±11	48.8±11	67.5±10	62.5±11	71.3±10	50.0±11	60.0±11	42.5±11	55.8	
	SEN	66.4±9	78.5±8	72.0±9	73.8±8	71.0±9	59.8±9	62.6±9	64.5±9	72.9±8	69.2±9	82.2±7	70.3	
	AUC	61.0±7	64.0±7	63.0±7	63.0±7	60.0±7	64.0±7	63.0±7	**68.0±7**	61.0±7	65.0±7	62.0±7	63.1	
	MCC	0.21	0.30	0.27	0.26	0.20	0.27	0.25	0.35	0.23	0.29	0.27	-	
	F1	0.66	0.73	0.70	0.70	0.68	0.65	0.66	0.69	0.69	0.69	0.73	-	
	MeN	1.00	0.03	0.46	0.18	0.24	0.04	0.23	0.05	0.19	0.90	0	-	
	χ^2	0.05	0	0.01	0	0.01	0.03	0.05	0	0	0.01	0	-	
25%	ACC	65.2±7	66.3±7	69.5±7	65.2±7	66.3±7	64.7±7	64.2±7	70.6±7	64.2±7	68.5±7	63.6±7	66.2	25
	SPE	55.0±11	51.3±11	57.5±11	63.8±11	53.8±11	51.3±11	52.5±11	70.0±10	58.8±11	58.8±11	50.0±11	56.6	
	SEN	72.9±8	77.6±8	78.5±8	66.4±9	75.7±8	74.8±8	72.9±8	71.0±9	68.2±9	75.7±8	73.8±8	73.4	
	AUC	64.0±7	64.0±7	68.0±7	65.0±7	65.0±7	63.0±7	63.0±7	**71.0±7**	63.0±7	67.0±7	62.0±7	**65.0**	
	MCC	0.28	0.30	0.37	0.30	0.30	0.27	0.26	0.41	0.27	0.35	0.25	-	
	F1	0.71	0.72	0.75	0.69	0.72	0.71	0.70	0.73	0.69	0.73	0.70	-	
	MeN	0.39	0.06	0.15	0.39	0.17	0.14	0.27	0.35	0.90	0.36	0.15	-	
	χ^2	0	0	0	0.01	0	0	0	0	0.02	0	0	-	
AVG	AUC	62.3	63.3	64.0	63.0	62.0	62.7	62.7	**68.0**	62.0	65.0	61.3	-	

Table A9. Representation of diagnostic performance for the external validation of cohort 2 using core sub-regions under different models. The training set consisted of cohort 1, which included 187 samples (107 high-grade and 80 low-grade), while cohort 2, comprising 204 samples (77 high-grade and 127 low-grade), was utilised for testing/external validation set.

COHORT 2 CORE N=204 EXTERNAL VALIDATION													
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	72.1±6	72.1±6	68.1±6	76.5±6	67.7±6	63.2±7	74.5±6	74.0±6	72.6±6	73.0±6	74.0±6	71.6	34
	SPE	70.9±8	70.9±8	67.7±8	92.1±5	67.7±8	56.7±9	84.3±6	83.5±6	78.0±7	74.0±8	89.0±5	75.9	
	SEN	74.0±10	74.0±10	68.8±10	50.7±11	67.5±10	74.0±10	58.4±11	58.4±11	63.6±11	71.4±10	49.4±11	64.6	
	AUC	72.0±6	72.0±6	68.0±6	71.0±6	68.0±6	65.0±7	71.0±6	71.0±6	71.0±6	**73.0±6**	69.0±6	70.1	
	MCC	0.45	0.44	0.36	0.49	0.34	0.30	0.44	0.43	0.42	0.44	0.43	-	
	F1	0.67	0.67	0.62	0.62	0.61	0.60	0.63	0.63	0.64	0.67	0.59	-	
	MeN	0.02	0.02	0.03	0	0.05	0	0.10	0.13	1.00	0.14	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
50%	ACC	78.9±6	77.9±6	71.6±6	78.4±6	73.0±6	74.0±6	76.5±6	78.9±6	72.6±6	76.0±6	77.9±6	76.0	24
	SPE	83.5±6	79.5±7	70.1±8	89.8±5	71.7±8	80.3±7	85.8±6	89.8±5	74.8±8	77.2±7	91.3±5	81.3	
	SEN	71.4±10	75.3±10	74.0±10	59.7±11	75.3±10	63.6±10	61.0±11	61.0±11	68.8±10	74.0±10	55.8±11	67.3	
	AUC	**77.0±6**	77.0±6	72.0±6	75.0±6	73.0±6	72.0±6	73.0±6	75.0±6	72.0±6	76.0±6	74.0±6	**74.2**	
	MCC	0.55	0.54	0.43	0.53	0.46	0.44	0.49	0.54	0.43	0.50	0.52	-	
	F1	0.72	0.72	0.66	0.68	0.68	0.65	0.66	0.69	0.65	0.70	0.66	-	
	MeN	0.88	0.30	0.02	0.01	0.02	0.68	0.08	0.01	0.29	0.20	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
25%	ACC	72.1±6	74.5±6	72.6±6	76.3±6	73.0±6	70.1±6	73.0±6	72.6±6	68.14±6	74.5±6	75.0±6	72.9	22
	SPE	78.7±7	77.2±7	76.4±7	90.6±5	82.7±7	78.0±7	88.2±6	81.9±7	70.1±8	82.7±7	86.6±6	81.2	
	SEN	61.0±11	70.1±10	66.2±11	53.3±11	57.1±11	57.1±11	48.1±11	57.1±11	64.9±11	61.0±11	55.8±11	59.3	
	AUC	70.0±6	**74.0±6**	71.0±6	72.0±6	70.0±6	68.0±6	68.0±6	70.0±6	68.0±6	72.0±6	71.0±6	70.4	
	MCC	0.40	0.47	0.42	0.48	0.41	0.36	0.40	0.40	0.34	0.45	0.45	-	
	F1	0.62	0.67	0.65	0.63	0.62	0.59	0.57	0.61	0.61	0.64	0.63	-	
	MeN	0.69	0.41	0.59	0	0.14	0.52	0	0.18	0.17	0.27	0.02	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
AVG	AUC	73.0	**74.3**	70.3	72.8	70.3	66.3	70.7	72.0	70.3	73.7	71.3	-	
100%	ACC	75.0±6	77.0±6	71.6±6	75.6±6	71.1±6	76.5±6	75.0±6	74.5±6	73.5±6	72.1±6	73.5±6	74.1	32
	SPE	79.5±7	79.5±7	81.9±7	88.2±6	73.2±8	90.6±5	81.1±7	81.1±7	77.2±7	74.0±8	89.0±5	81.4	
	SEN	67.5±10	72.7±10	54.6±11	54.6±11	67.5±10	53.3±11	64.9±11	63.6±11	67.5±10	68.8±10	48.1±11	62.1	
	AUC	74.0±6	**76.0±6**	68.0±6	71.0±6	70.0±6	72.0±6	73.0±6	72.0±6	72.0±6	71.0±6	69.0±6	**71.6**	
	MCC	0.47	0.52	0.38	0.46	0.4	0.48	0.46	0.45	0.44	0.42	0.41	-	
	F1	0.67	0.70	0.59	0.63	0.64	0.63	0.66	0.65	0.66	0.65	0.58	-	
	MeN	0.89	0.47	0.12	0	0.24	0	0.67	0.58	0.59	0.23	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	

 doi:10.20944/preprints202312.1379.v5

Table A10. Representation of diagnostic performance for the external validation of cohort 2 using periphery sub-regions under different models. Cohort 1, comprising 187 samples (107 high-grade and 80 low-grade), was utilised as the training set, while Cohort 2, consisting of 204 samples (77 high-grade and 127 low-grade), served for testing/external validation set.

COHORT 2 PERIPHERY N=204 EXTERNAL VALIDATION													
REGION		SVM	RF	XGB	NB	MLP	LSTM	LR	QDA	LGB	CB	ADB	AVG	FS
75%	ACC	76.0±6	75.0±6	67.7±6	72.1±6	70.6±6	74.0±6	75.5±6	70.1±6	68.1±6	68.1±6	72.1±6	71.8	
	SPE	80.3±7	74.8±8	74.8±8	78.7±7	75.6±7	87.4±6	87.4±6	68.5±8	77.2±7	70.1±8	88.2±6	78.5	
	SEN	68.8±10	75.3±10	55.8±11	61.0±11	62.3±11	52.0±11	55.8±11	72.7±10	53.3±11	64.9±11	45.5±11	60.7	
	AUC	75.0±6	**75.0±6**	65.0±7	70.0±6	69.0±6	70.0±6	72.0±6	71.0±6	65.0±7	68.0±6	67.0±7	69.7	28
	MCC	0.49	0.49	0.31	0.40	0.38	0.43	0.46	0.40	0.31	0.34	0.38	-	
	F1	0.68	0.69	0.57	0.62	0.62	0.60	0.63	0.65	0.56	0.61	0.55	-	
	McN	0.89	0.07	0.81	0.69	0.80	0	0.01	0.01	0.39	0.17	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
50%	ACC	74.0±6	75.5±6	66.7±6	71.6±6	74.5±6	74.5±6	74.0±6	75.5±6	68.1±6	70.6±6	75.5±6	72.8	
	SPE	76.4±6	74.8±8	67.7±8	82.7±7	85.0±6	85.0±6	79.5±7	86.6±6	75.6±7	77.2±7	94.5±4	80.5	
	SEN	70.1±11	76.6±9	64.9±11	53.3±11	57.1±11	57.1±11	64.9±11	57.1±11	55.8±11	59.7±11	44.0±11	60.1	
	AUC	73.0±6	**76.0±6**	66.0±7	68.0±6	71.0±6	71.0±6	72.0±6	72.0±6	66.0±7	68.0±6	69.0±6	70.2	28
	MCC	0.46	0.50	0.32	0.38	0.44	0.44	0.45	0.46	0.32	0.37	0.47	-	
	F1	0.67	0.70	0.60	0.59	0.63	0.63	0.65	0.64	0.57	0.61	0.58	-	
	McN	0.34	0.05	0.09	0.07	0.05	0.05	0.89	0.02	0.71	0.80	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
25%	ACC	75.5±6	75.5±6	67.2±6	74.5±6	69.1±6	71.6±6	77.5±6	76.0±6	71.6±6	72.1±6	75.0±6	73.2	
	SPE	77.2±7	75.6±7	74.8±8	88.2±6	67.7±8	81.1±7	88.2±6	85.0±6	77.2±7	74.02±8	91.3±5	80.0	
	SEN	72.7±10	75.3±10	54.6±11	52.0±11	71.4±10	55.8±11	59.7±11	61.0±11	62.3±11	68.8±10	48.1±11	62.0	
	AUC	75.0±6	**75.0±6**	65.0±7	70.0±6	70.0±6	68.0±6	74.0±6	73.0±6	70.0±6	71.0±6	70.0±6	**71.0**	24
	MCC	0.49	0.50	0.30	0.44	0.38	0.38	0.51	0.48	0.40	0.42	0.45	-	
	F1	0.69	0.70	0.56	0.61	0.64	0.60	0.67	0.66	0.62	0.65	0.59	-	
	McN	0.26	0.09	0.71	0	0.02	0.19	0.02	0.12	1.00	0.23	0	-	
	χ^2	0	0	0	0	0	0	0	0	0	0	0	-	
AVG	AUC	74.3	**75.3**	65.3	69.3	70.0	69.7	72.7	72.0	67.0	69.0	68.7	-	

Table A11. Diagnostic performance for best-performing regions in cohort 4. 159 samples from cohort 1 were allocated to the training set, comprising 96 high-grade and 63 low-grade samples. Additionally, cohort 4, with 28 samples, was designated for the testing/internal validation set, including 11 high-grade and 17 low-grade samples. On the other hand, Cohort 2, consisting of 77 high-grade and 127 low-grade samples, served as the training set, while cohort 4 was reserved for the testing/external validation set.

COHORT 4						
INTERNAL VALIDATION				EXTERNAL VALIDATION		
REGION		QDA	CB	REGION	QDA	CB
75% CORE	ACC	82.1±14	85.7±13	50% CORE	71.4±17	75.0±16
	SPE	94.1±9	94.1±9		82.4±18	70.6±22
	SEN	63.6±28	72.7±26		54.6±29	81.8±20
	AUC	79.0±15	83.0±14		68.0±17	76.0±16
	MCC	0.62	0.70		0.39	0.51
	F1	0.74	0.80		0.60	0.72
	McN	0.18	0.32		0.48	0.26
	χ^2	0.03	0.02		0.23	0.20
50% CORE	ACC	82.1±14	92.9±8	50% PERIPHERY	67.9±17	78.6±15
	SPE	70.6±22	94.1±9		64.7±23	82.4±18
	SEN	100.0±14	90.9±13		72.7±26	72.7±26
	AUC	85.0±13	93.0±10		69.0±17	78.0±15
	MCC	0.70	0.85		0.37	0.55
	F1	0.81	0.91		0.64	0.73
	McN	0.03	1.00		0.32	1.00
	χ^2	0.02	0		0.45	0.13
25% PERIPHERY	ACC	75.0±16	96.4±5	25% PERIPHERY	75.0±16	82.1±14
	SPE	64.7±23	100.0±12		76.5±20	88.2±14
	SEN	90.9±13	90.9±13		72.7±26	72.7±26
	AUC	78.0±15	**95.0±8**		75.0±16	**80.0±15**
	MCC	0.55	0.93		0.49	0.62
	F1	0.74	0.95		0.70	0.76
	McN	0.06	0.32		0.71	0.65
	χ^2	0.11	0		0.23	0.06
AVG	AUC	80.7	**90.3**		70.7	**78.0**
100%	ACC	85.7±13	92.9±8	100%	78.6±15	78.6±15
	SPE	94.1±9	94.1±9		88.2±14	76.5±20
	SEN	72.7±26	90.0±14		63.6±28	81.8±20
	AUC	83.0±14	93.0±10		76.0±16	79.0±15
	MCC	0.70	0.85		0.54	0.57
	F1	0.80	0.91		0.70	0.75
	McN	0.32	1.00		0.41	0.41
	χ^2	0.02	0		0.09	0.13

 doi:10.20944/preprints202312.1379.v5

Name	Kidney Side	Age Y/M	Gender	Grade-Surgery-Histology	Grade-Biopsy-Histology	Grade of Biopsy Compared to Nephrectomy
AA014	L	58.2	M	2 ISUP-PN	Biopsy (Low Grade)	Correct
AA015	R	60.3	F	2 ISUP-RN	Biopsy (Undetermenate)	Not possible
AA054	L	41	M	2 ISUP-PN	Biopsy (Low Grade)	Correct
AA058	L	51.7	F	2 ISUP-PN	Biopsy (Low Grade)	Correct
AA075	L	71.7	M	3 ISUP-PN	**Biopsy (Low Grade)**	Wrong
AA076	L	54.11	M	2 ISUP-PN	Biopsy (No Grade)	Not possible
AA080	R	61	M	2 ISUP-PN	Biopsy (Undetermative)	Not possible
AA083	L	71.3	M	3 ISUP-PN	**Biopsy (Low Grade)**	Wrong
AA085	L	65.8	F	2 ISUP-PN	Biopsy (Low Grade)	Correct
AA088	L	40.2	M	3 ISUP-PN	**Biopsy (Low Grade)**	Wrong
AA095	L	67.1	M	2 ISUP-PN	Biopsy (No Grade)	Not possible
AA160	RT	67	M	2 ISUP-RN	Biopsy (Low Grade)	Correct
AA176	LT	81.5	M	3 ISUP-RN	Biopsy (No Grade)	Not possible
AA191	RT	61	M	2 ISUP-PN	Biopsy (Undeterminate)	Not possible
AA213	LT	67.1	F	4 ISUP-RN	Biopsy (High Grade)	Correct
AA219	LT	71.4	M	3 ISUP-PN	**Biopsy (Low Grade)**	Wrong
AA226	RT	67	M	2 ISUP-RN	Biopsy (Low Grade)	Correct
AA235	LT	68.5	F	2 ISUP-PN	Biopsy (Undetermative)	Not possible
AA254	LT	55.2	M	2 ISUP-RN	Biopsy (Low Grade)	Correct
AA270	LT	66.11	M	1 ISUP-RN	Biopsy (Low Grade)	Correct
AA271	RT	62.8	F	3 ISUP-RN	**Biopsy (ISUP 2)**	Wrong
AA285	LT	71.11	M	3 ISUP-PN	**Biopsy (Low Grade)**	Wrong
AA293	LT	50	M	High Grade-RN	Biopsy (Undetermative)	Not possible
AA319	RT	71.8	F	3 ISUP-RN	Biopsy (Undetermative)	Not possible
AA340	LT	61.11	M	2 ISUP-RN	Biopsy (Low Grade)	Correct
AA364	RT	75.7	M	3 ISUP-RN	Biopsy (No Grade)	Not possible
AA365	LT	38.6	M	2 ISUP-PN	Biopsy (No Grade)	Not possible
AA373	LT	67	M	Low Grade-PN	Biopsy (Undetermative)	Not possible

Figure A1. PN/RN versus biopsy grade histology.

Nephrectomy	Biopsy	Count	%Accuracy
High	Correct	1	9.09%
11	Wrong	10	
Low	Wrong	8	52.94%
17	Correct	9	

Figure A2. Summary of biopsy grade histology.

 doi:10.20944/preprints202312.1379.v5

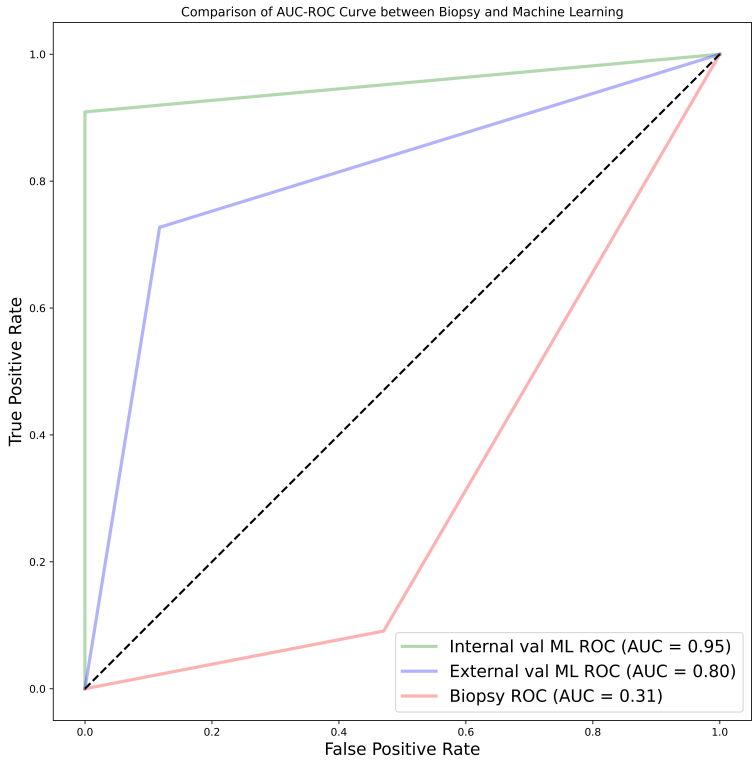


Figure A3. AUC-ROC curves representing the comparison between biopsy and machine learning.

References

1. Delahunt, B.; Eble, J.N.; Egevad, L.; Samaratunga, H. Grading of renal cell carcinoma. *Histopathology* **2019**, *74*, 4–17.
2. Delahunt, B.; Sika-Paotonu, D.; Bethwaite, P.B.; Jordan, T.W.; Magi-Galluzzi, C.; Zhou, M.; Samaratunga, H.; Srigley, J.R. Grading of clear cell renal cell carcinoma should be based on nucleolar prominence. *The American journal of surgical pathology* **2011**, *35*, 1134–1139.
3. Fuhrman, S.A.; Lasky, L.C.; Limas, C. Prognostic significance of morphologic parameters in renal cell carcinoma. *The American journal of surgical pathology* **1982**, *6*, 655–664.
4. Dagher, J.; Delahunt, B.; Rioux-Leclercq, N.; Egevad, L.; Srigley, J.R.; Coughlin, G.; Dunglinson, N.; Gianduzzo, T.; Kua, B.; Malone, G.; others. Clear cell renal cell carcinoma: validation of World Health Organization/International Society of Urological Pathology grading. *Histopathology* **2017**, *71*, 918–925.
5. Moch, H.; Cubilla, A.L.; Humphrey, P.A.; Reuter, V.E.; Ulbright, T.M. The 2016 WHO classification of tumours of the urinary system and male genital organs—part A: renal, penile, and testicular tumours. *European urology* **2016**, *70*, 93–105.
6. Coy, H.; Douek, M.; Young, J.; Brown, M.S.; Goldin, J.; Sayre, J.; Raman, S. Differentiation of low grade from high grade clear cell renal cell carcinoma neoplasms using a CAD algorithm on four-phase CT., 2016.
7. Jeon, H.G.; Seo, S.I.; Jeong, B.C.; Jeon, S.S.; Lee, H.M.; Choi, H.Y.; Song, C.; Hong, J.H.; Kim, C.S.; Ahn, H.; others. Percutaneous kidney biopsy for a small renal mass: a critical appraisal of results. *The Journal of urology* **2016**, *195*, 568–573.
8. Delahunt, B.; Cheville, J.C.; Martignoni, G.; Humphrey, P.A.; Magi-Galluzzi, C.; McKenney, J.; Egevad, L.; Algaba, F.; Moch, H.; Grignon, D.J.; others. The International Society of Urological Pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters. *The American journal of surgical pathology* **2013**, *37*, 1490–1504.
9. Remzi, M.; Marberger, M. Renal tumor biopsies for evaluation of small renal tumors: why, in whom, and how? *European urology* **2009**, *55*, 359–367.
10. Lane, B.R.; Samplaski, M.K.; Herts, B.R.; Zhou, M.; Novick, A.C.; Campbell, S.C. Renal mass biopsy—a renaissance? *The Journal of urology* **2008**, *179*, 20–27.

 doi:10.20944/preprints202312.1379.v5

11. Dhaun, N.; Bellamy, C.O.; Cattran, D.C.; Kluth, D.C. Utility of renal biopsy in the clinical management of renal disease. *Kidney international* **2014**, *85*, 1039–1048.
12. Andersen, M.; Norus, T. Tumor seeding with renal cell carcinoma after renal biopsy. *Urology Case Reports* **2016**, *9*, 43–44.
13. Volpe, A.; Mattar, K.; Finelli, A.; Kachura, J.R.; Evans, A.J.; Geddie, W.R.; Jewett, M.A. Contemporary results of percutaneous biopsy of 100 small renal masses: a single center experience. *The Journal of urology* **2008**, *180*, 2333–2337.
14. Wu, J.; Gong, G.; Cui, Y.; Li, R. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging* **2016**, *44*, 1107–1115.
15. Synnott, N.C.; Poeta, M.L.; Costantini, M.; Pfeiffer, R.M.; Li, M.; Golubeva, Y.; Lawrence, S.; Mutreja, K.; Amoreo, C.; Dabrowska, M.; others. Characterizing the tumor microenvironment in rare renal cancer histological types. *The Journal of Pathology: Clinical Research* **2022**, *8*, 88–98.
16. Wu, J.; Aguilera, T.; Shultz, D.; Gudur, M.; Rubin, D.L.; Loo Jr, B.W.; Diehn, M.; Li, R. Early-stage non–small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. *Radiology* **2016**, *281*, 270–278.
17. Serganova, I.; Doubrovin, M.; Vider, J.; Ponomarev, V.; Soghomonyan, S.; Beresten, T.; Ageyeva, L.; Serganov, A.; Cai, S.; Balatoni, J.; others. Molecular imaging of temporal dynamics and spatial heterogeneity of hypoxia-inducible factor-1 signal transduction activity in tumors in living mice. *Cancer Research* **2004**, *64*, 6101–6108.
18. Qu, J.Y.; Jiang, H.; Song, X.H.; Wu, J.K.; Ma, H. Four-phase computed tomography helps differentiation of renal oncocytoma with central hypodense areas from clear cell renal cell carcinoma. *Diagn Interv Radiol* **2022**.
19. Qu, J.; Zhang, Q.; Song, X.; Jiang, H.; Ma, H.; Li, W.; Wang, X. CT differentiation of the oncocytoma and renal cell carcinoma based on peripheral tumor parenchyma and central hypodense area characterisation. *BMC Medical Imaging* **2023**, *23*, 16.
20. Teifke, A.; Behr, O.; Schmidt, M.; Victor, A.; Vomweg, T.W.; Thelen, M.; Lehr, H.A. Dynamic MR imaging of breast lesions: correlation with microvessel distribution pattern and histologic characteristics of prognosis. *Radiology* **2006**, *239*, 351–360.
21. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577.
22. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; Van Stiphout, R.G.; Granton, P.; Zegers, C.M.; Gillies, R.; Boellard, R.; Dekker, A.; others. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* **2012**, *48*, 441–446.
23. Alhussaini, A.J.; Steele, J.D.; Nabi, G. Comparative Analysis for the Distinction of Chromophobe Renal Cell Carcinoma from Renal Oncocytoma in Computed Tomography Imaging Using Machine Learning Radiomics Analysis. *Cancers* **2022**, *14*, 3609.
24. Demirjian, N.L.; Varghese, B.A.; Cen, S.Y.; Hwang, D.H.; Aron, M.; Siddiqui, I.; Fields, B.K.; Lei, X.; Yap, F.Y.; Rivas, M.; others. CT-based radiomics stratification of tumor grade and TNM stage of clear cell renal cell carcinoma. *European Radiology* **2022**, pp. 1–12.
25. Nazari, M.; Shiri, I.; Zaidi, H. Radiomics-based machine learning model to predict risk of death within 5-years in clear cell renal cell carcinoma patients. *Computers in biology and medicine* **2021**, *129*, 104135.
26. Lin, M.; Wynne, J.F.; Zhou, B.; Wang, T.; Lei, Y.; Curran, W.J.; Liu, T.; Yang, X. Artificial intelligence in tumor subregion analysis based on medical imaging: A review. *Journal of Applied Clinical Medical Physics* **2021**, *22*, 10–26.
27. Arteaga-Arteaga, H.B.; Candamil-Cortés, M.S.; Breaux, B.; Guillen-Rondon, P.; Orozco-Arias, S.; Tabares-Soto, R. Machine learning applications on intratumoral heterogeneity in glioblastoma using single-cell RNA sequencing data. *Briefings in Functional Genomics* **2023**, p. elad002.
28. Pan, Z.; Men, K.; Liang, B.; Song, Z.; Wu, R.; Dai, J. A subregion-based prediction model for local–regional recurrence risk in head and neck squamous cell carcinoma. *Radiotherapy and Oncology* **2023**, *184*, 109684.
29. Lu, H.; Yin, J. Texture analysis of breast DCE-MRI based on intratumoral subregions for predicting HER2 2+ status. *Frontiers in Oncology* **2020**, *10*, 543.

 doi:10.20944/preprints202312.1379.v5

30. Heller, N.; Isensee, F.; Maier-Hein, K.H.; Hou, X.; Xie, C.; Li, F.; Nan, Y.; Mu, G.; Lin, Z.; Han, M.; others. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical image analysis* **2021**, *67*, 101821.
31. Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; others. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445* **2019**.
32. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; others. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* **2013**, *26*, 1045–1057.
33. Shu, J.; Wen, D.; Xi, Y.; Xia, Y.; Cai, Z.; Xu, W.; Meng, X.; Liu, B.; Yin, H. Clear cell renal cell carcinoma: Machine learning-based computed tomography radiomics analysis for the prediction of WHO/ISUP grade. *European journal of radiology* **2019**, *121*, 108738.
34. Vobugari, N.; Raja, V.; Sethi, U.; Gandhi, K.; Raja, K.; Surani, S.R. Advancements in oncology with artificial intelligence—A review article. *Cancers* **2022**, *14*, 1349.
35. Nazari, M.; Shiri, I.; Hajianfar, G.; Oveisi, N.; Abdollahi, H.; Deevband, M.R.; Oveisi, M.; Zaidi, H. Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning. *La radiologia medica* **2020**, *125*, 754–762.
36. Sun, X.; Liu, L.; Xu, K.; Li, W.; Huo, Z.; Liu, H.; Shen, T.; Pan, F.; Jiang, Y.; Zhang, M. Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. *Medicine* **2019**, *98*.
37. Goldman, L.W. Principles of CT: radiation dose and image quality. *Journal of nuclear medicine technology* **2007**, *35*, 213–225.
38. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer research* **2017**, *77*, e104–e107.
39. Ganeshan, B.; Goh, V.; Mandeville, H.C.; Ng, Q.S.; Hoskin, P.J.; Miles, K.A. Non–small cell lung cancer: histopathologic correlates for texture parameters at CT. *Radiology* **2013**, *266*, 326–336.
40. Debie, E.; Shafi, K. Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. *Pattern Analysis and Applications* **2019**, *22*, 519–536.
41. Brownlee, J. How to Avoid Data Leakage When Performing Data Preparation. https://machinelearningmastery.com/data-preparation-without-data-leakage/ accessed (27 April 2023).
42. Park, J.E.; Kim, D.; Kim, H.S.; Park, S.Y.; Kim, J.Y.; Cho, S.J.; Shin, J.H.; Kim, J.H. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. *European radiology* **2020**, *30*, 523–536.
43. Karagöz, A.; Guvenis, A. Robust whole-tumour 3D volumetric CT-based radiomics approach for predicting the WHO/ISUP grade of a ccRCC tumour. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **2023**, *11*, 665–677.
44. Zhao, Y.; Fu, X.; Lopez, J.I.; Rowan, A.; Au, L.; Fendler, A.; Hazell, S.; Xu, H.; Horswell, S.; Shepherd, S.T.; others. Selection of metastasis competent subclones in the tumour interior. *Nature ecology & evolution* **2021**, *5*, 1033–1045.
45. He, X.; Wei, Y.; Zhang, H.; Zhang, T.; Yuan, F.; Huang, Z.; Han, F.; Song, B. 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* **2020**, *27*, 157–168.
46. Xv, Y.; Lv, F.; Guo, H.; Zhou, X.; Tan, H.; Xiao, M.; Zheng, Y. Machine learning-based CT radiomics approach for predicting WHO/ISUP nuclear grade of clear cell renal cell carcinoma: an exploratory and comparative study. *Insights Into Imaging* **2021**, *12*, 1–14.
47. Cui, E.; Li, Z.; Ma, C.; Li, Q.; Lei, Y.; Lan, Y.; Yu, J.; Zhou, Z.; Li, R.; Long, W.; others. Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. *European Radiology* **2020**, *30*, 2912–2921.
48. Wang, R.; Hu, Z.; Shen, X.; Wang, Q.; Zhang, L.; Wang, M.; Feng, Z.; Chen, F. Computed tomography-based radiomics model for predicting the WHO/ISUP grade of clear cell renal cell carcinoma preoperatively: a multicenter study. *Frontiers in Oncology* **2021**, *11*, 543854.

 doi:10.20944/preprints202312.1379.v5

49. Moldovanu, C.G.; Boca, B.; Lebovici, A.; Tamas-Szora, A.; Feier, D.S.; Crisan, N.; Andras, I.; Buruian, M.M. Preoperative predicting the WHO/ISUP nuclear grade of clear cell renal cell carcinoma by computed tomography-based radiomics features. *Journal of Personalized Medicine* **2021**, *11*, 8.
50. Yi, X.; Xiao, Q.; Zeng, F.; Yin, H.; Li, Z.; Qian, C.; Wang, C.; Lei, G.; Xu, Q.; Li, C.; others. Computed tomography radiomics for predicting pathological grade of renal cell carcinoma. *Frontiers in oncology* **2021**, *10*, 570396.
51. Paterson, C.; Ghaemi, J.; Alashkham, A.; Biyani, C.S.; Coles, B.; Baker, L.; Szewczyk-Bieda, M.; Nabi, G. Diagnostic accuracy of image-guided biopsies in small (< 4 cm) renal masses with implications for active surveillance: a systematic review of the evidence. *The British Journal of Radiology* **2018**, *91*, 20170761.
52. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl* **2013**, *3*.
53. Shi, L.; Campbell, G.; Jones, W.; Campagne, F.; Wen, Z.; Walker, S.; Su, Z.; Chu, T.; Goodsaid, F.; Pusztai, L.; others. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology* **2010**, *28*, 827–838.
54. Brown, J. Classifiers and their metrics quantified. *Molecular informatics* **2018**, *37*, 1700127.
55. Alzahrani, S.; Al-Bander, B.; Al-Nuaimy, W. A comprehensive evaluation and benchmarking of convolutional neural networks for melanoma diagnosis. *Cancers* **2021**, *13*, 4494.
56. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **2015**, *10*, e0118432.
57. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* **2020**, *21*, 1–13.
58. Muglia, V.F.; Prando, A. Carcinoma de células renais: classificação histológica e correlação com métodos de imagem. *Radiologia Brasileira* **2015**, *48*, 166–174.
59. Delahunt, B.; Egevad, L.; Samaratunga, H.; Martignoni, G.; Nacey, J.N.; Srigley, J.R. Gleason and Fuhrman no longer make the grade. *Histopathology* **2016**, *68*, 475–481.
60. Samaratunga, H.; Gianduzzo, T.; Delahunt, B. The ISUP system of staging, grading and classification of renal cell neoplasia. *Journal of kidney cancer and VHL* **2014**, *1*, 26.
61. Delahunt, B. Advances and controversies in grading and staging of renal cell carcinoma. *Modern Pathology* **2009**, *22*, S24–S36.
62. Lang, H.; Lindner, V.; de Fromont, M.; Molinié, V.; Letourneux, H.; Meyer, N.; Martin, M.; Jacqmin, D. Multicenter determination of optimal interobserver agreement using the Fuhrman grading system for renal cell carcinoma: assessment of 241 patients with> 15-year follow-up. *Cancer* **2005**, *103*, 625–629.
63. Chartrand, G.; Cheng, P.M.; Vorontsov, E.; Drozdzal, M.; Turcotte, S.; Pal, C.J.; Kadoury, S.; Tang, A. Deep learning: a primer for radiologists. *Radiographics* **2017**, *37*, 2113–2131.
64. Glaßer, S.; Niemann, U.; Preim, B.; Spiliopoulou, M. Can we distinguish between benign and malignant breast tumors in DCE-MRI by studying a tumor's most suspect region only? Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. IEEE, 2013, pp. 77–82.
65. Karahaliou, A.; Vassiou, K.; Arikidis, N.; Skiadopoulos, S.; Kanavou, T.; Costaridou, L. Assessing heterogeneity of lesion enhancement kinetics in dynamic contrast-enhanced MRI for breast cancer diagnosis. *The British journal of radiology* **2010**, *83*, 296–309.
66. Milenković, J.; Hertl, K.; Košir, A.; Žibert, J.; Tasič, J.F. Characterization of spatiotemporal changes for the classification of dynamic contrast-enhanced magnetic-resonance breast lesions. *Artificial intelligence in medicine* **2013**, *58*, 101–114.
67. Agner, S.; Rosen, M.; Englander, S.; Sobers, D.; Thomas, K.; Tomaszewski, J.; Feldman, M.; Ganesan, S.; Schnall, M.; Madabhushi, A. Distinguishing molecular subtypes of breast cancer based on computer-aided diagnosis of dce-mri. International Society for Magnetic Resonance in Medicine Annual Meeting, 2010, Vol. 2490.
68. Chaudhury, B.; Zhou, M.; Goldgof, D.B.; Hall, L.O.; Gatenby, R.A.; Gillies, R.J.; Drukteinis, J.S. Using features from tumor subregions of breast dce-mri for estrogen receptor status prediction. 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2014, pp. 2624–2629.
69. Mahrooghy, M.; Ashraf, A.B.; Daye, D.; Mies, C.; Feldman, M.; Rosen, M.; Kontos, D. Heterogeneity wavelet kinetics from DCE-MRI for classifying gene expression based breast cancer recurrence risk. Medical Image

Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16. Springer, 2013, pp. 295–302.
70. Zhang, L.; Wang, Y.; Peng, Z.; Weng, Y.; Fang, Z.; Xiao, F.; Zhang, C.; Fan, Z.; Huang, K.; Zhu, Y.; others. The progress of multimodal imaging combination and subregion based radiomics research of cancers. *International Journal of Biological Sciences* **2022**, *18*, 3458.
71. Meng, X.; Shu, J.; Xia, Y.; Yang, R. A CT-based radiomics approach for the differential diagnosis of sarcomatoid and clear cell renal cell carcinoma. *BioMed Research International* **2020**, *2020*.
72. Takahashi, N.; Leng, S.; Kitajima, K.; Gomez-Cardona, D.; Thapa, P.; Carter, R.E.; Leibovich, B.C.; Sasiwimonphan, K.; Sasaguri, K.; Kawashima, A. Small (< 4 cm) renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma using unenhanced and contrast-enhanced CT. *American Journal of Roentgenology* **2015**, *205*, 1194–1202.
73. Varga, T.V.; Niss, K.; Estampador, A.C.; Collin, C.B.; Moseley, P.L. Association is not prediction: a landscape of confused reporting in diabetes–a systematic review. *Diabetes research and clinical practice* **2020**, *170*, 108497.
74. Faber, J.; Fonseca, L.M. How sample size influences research outcomes. *Dental press journal of orthodontics* **2014**, *19*, 27–29.
75. Xu, L.; Yang, C.; Zhang, F.; Cheng, X.; Wei, Y.; Fan, S.; Liu, M.; He, X.; Deng, J.; Xie, T.; others. Deep learning using CT images to grade clear cell renal cell carcinoma: development and validation of a prediction model. *Cancers* **2022**, *14*, 2574.
76. Wen-Zhi, G.; Tai, T.; Zhixin, F.; Huanyu, L.; Yanqing, G.; Yuexian, G.; Xuesong, L. Prediction of pathological staging and grading of renal clear cell carcinoma based on deep learning algorithms. *Journal of International Medical Research* **2022**, *50*, 03000605221135163.
77. Lin, F.; Ma, C.; Xu, J.; Lei, Y.; Li, Q.; Lan, Y.; Sun, M.; Long, W.; Cui, E. A CT-based deep learning model for predicting the nuclear grade of clear cell renal cell carcinoma. *European Journal of Radiology* **2020**, *129*, 109079.
78. Lechevallier, E.; André, M.; Barriol, D.; Daniel, L.; Eghazarian, C.; De Fromont, M.; Rossi, D.; Coulange, C. Fine-needle percutaneous biopsy of renal masses with helical CT guidance. *Radiology* **2000**, *216*, 506–510.
79. Volpe, A.; Finelli, A.; Gill, I.S.; Jewett, M.A.; Martignoni, G.; Polascik, T.J.; Remzi, M.; Uzzo, R.G. Rationale for percutaneous biopsy and histologic characterisation of renal tumours. *European urology* **2012**, *62*, 491–504.
80. Blumenfeld, A.J.; Guru, K.; Fuchs, G.J.; Kim, H.L. Percutaneous biopsy of renal cell carcinoma underestimates nuclear grade. *Urology* **2010**, *76*, 610–613.
81. Lebret, T.; Poulain, J.E.; Molinie, V.; Herve, J.M.; Denoux, Y.; Guth, A.; Scherrer, A.; Botto, H. Percutaneous core biopsy for renal masses: indications, accuracy and results. *The Journal of urology* **2007**, *178*, 1184–1188.
82. Millet, I.; Curros, F.; Serre, I.; Taourel, P.; Thuret, R. Can renal biopsy accurately predict histological subtype and Fuhrman grade of renal cell carcinoma? *The Journal of urology* **2012**, *188*, 1690–1694.
83. Neuzillet, Y.; Lechevallier, E.; Andre, M.; Daniel, L.; Coulange, C. Accuracy and clinical role of fine needle percutaneous biopsy with computerized tomography guidance of small (less than 4.0 cm) renal masses. *The Journal of urology* **2004**, *171*, 1802–1805.
84. Schmidbauer, J.; Remzi, M.; Memarsadeghi, M.; Haitel, A.; Klingler, H.C.; Katzenbeisser, D.; Wiener, H.; Marberger, M. Diagnostic accuracy of computed tomography-guided percutaneous biopsy of renal masses. *European urology* **2008**, *53*, 1003–1012.
85. Wunderlich, H.; Hindermann, W.; Mustafa, A.M.A.; Reichelt, O.; Junker, K.; SCHUBERT, J. The accuracy of 250 fine needle biopsies of renal tumors. *The Journal of urology* **2005**, *174*, 44–46.
86. Kunkle, D.A.; Egleston, B.L.; Uzzo, R.G. Excise, ablate or observe: the small renal mass dilemma—a meta-analysis and review. *The Journal of urology* **2008**, *179*, 1227–1234.