Original Article

Tabular prior-data fitted network in real-world CT radiomics: benign *vs.* malignant renal tumor classification

Tianzhu Liu[1#], Huanjun Wang[2#], Yan Guo[2], Yongsong Ye[3], Bei Weng[2], Xiaodan Li[4], Jun Chen[1], Shanghuang Xie[5], Guimian Zhong[6], Zhixuan Song[7], Lesheng Huang[1]

[1]Department of Radiology, Guangdong Provincial Hospital of Traditional Chinese Medicine, Zhuhai, China; [2]Department of Radiology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China; [3]Department of Radiology, Guangdong Provincial Hospital of Traditional Chinese Medicine, Guangzhou, China; [4]Department of Gynaecology, Guangdong Provincial Hospital of Traditional Chinese Medicine, Zhuhai, China; [5]Lab of Molecular Imaging and Medical Intelligence, Department of Radiology, Longgang Central Hospital of Shenzhen, Shenzhen Clinical Medical College, Guangzhou University of Chinese Medicine, Longgang Central Hospital of Shantou University Medical College, Shenzhen, China; [6]Department of Radiology, The Affiliated Panyu Central Hospital of Guangzhou Medical University, Guangzhou, China; [7]Clinical and Technical Support, Philips Healthcare, Guangzhou, China

Contributions: (I) Conception and design: T Liu, L Huang; (II) Administrative support: S Xie, J Chen, Y Guo, Y Ye; (III) Provision of study materials or patients: T Liu, Y Ye, B Weng, X Li, H Wang, G Zhong; (IV) Collection and assembly of data: T Liu, B Weng, Z Song; (V) Data analysis and interpretation: T Liu, Z Song, L Huang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Lesheng Huang, MSc. Department of Radiology, Guangdong Provincial Hospital of Traditional Chinese Medicine, No. 53 Jingle Road, Zhuhai 519000, China. Email: huanglesheng2008@126.com.

Background: Radiomics-based machine learning (ML) models can potentially distinguish between benign and malignant renal tumors on computed tomography (CT). Traditional algorithms necessitate intensive hyperparameter tuning and large datasets for optimization. This study assessed the novel tabular prior-data fitted network (TabPFN), a pre-trained transformer model for tabular data classification, and compared it with conventional algorithms using real-world CT radiomics datasets.

Methods: Two retrospective cohorts were analyzed: datasets A [207 cystic renal masses (CRMs): 92 benign, 115 malignant] and B [92 tumors: 41 renal oncocytomas (ROs), 51 chromophobe renal cell carcinomas (CRCCs)]. Radiomic features were extracted from three-phase CT images (unenhanced, corticomedullary, and nephrographic) using PyRadiomics. Seven algorithms [support vector machine (SVM), stochastic gradient descent (SGD), k-nearest neighbor (KNN), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and TabPFN] were evaluated via 10-fold cross-validation. Evaluation metrics included area under the curve (AUC), accuracy, sensitivity, specificity, and feature importance.

Results: In dataset A, all models achieved high AUCs (training: 0.935–1.000; validation: 0.800–0.946). TabPFN showed top-tier performance, particularly in nephrographic-phase analysis (validation AUC: 0.946). In dataset B, TabPFN demonstrated superior stability (validation AUC: 0.700–0.800), outperforming SVM and KNN while circumventing the convergence failures of SGD. On feature importance analysis, its dynamic weight allocation paralleled that of ensemble models (RF/XGBoost) without explicit hyperparameter optimization.

Conclusions: TabPFN demonstrates robust performance in differentiating renal tumors, particularly in small-scale, high-dimensional datasets. Its transformer-based architecture and pre-training on synthetic data eliminate manual parameter tuning, enhancing clinical applicability over conventional ML approaches.

Keywords: Tabular prior-data fitted network (TabPFN); machine learning (ML); renal tumor; real-world evaluation

Submitted May 13, 2025. Accepted for publication Sep 10, 2025. Published online Oct 22, 2025.
doi: 10.21037/qims-2025-1132
View this article at: https://dx.doi.org/10.21037/qims-2025-1132

Introduction

Radiomics, which facilitates the extraction of high-dimensional quantitative features from radiological images to assist in disease classification and prognostication, has emerged as a promising methodology in medical imaging analysis (1,2). Machine learning (ML) algorithms play an essential role in radiomics-based analysis. Traditional models such as support vector machines (SVMs), random forests (RFs) (3), and gradient boosting methods (4), although widely utilized (5,6), often require extensive hyperparameter tuning and large datasets to achieve optimal performance. This necessity limits their applicability in real-world clinical settings (7), where data availability may be constrained.

Introduced by Hollmann *et al.* in 2025, the tabular prior-data fitted network (TabPFN) algorithm (7) has shown promising capabilities in handling tabular datasets with minimal computational expense and has also eliminated the need for hyperparameter tuning. Unlike conventional ML models that depend on sample-specific learning (8), TabPFN employs a pre-trained transformer-based framework. This design allows for rapid and efficient learning across diverse datasets, significantly reducing training time while maintaining robust predictive accuracy. Importantly, it eliminates the need for hyperparameter tuning (a process that typically consumes above 80% of ML development time in radiomics studies) by dynamically adjusting attention weights during inference. This capability is transformative for tumor classification, as it offers robust performance even with fewer than 100 samples per class. These advantages suggest that TabPFN could be a powerful tool for clinical radiomics applications. Despite its theoretical benefits, the practical uses of TabPFN in real-world medical imaging and radiomics have remained largely unexplored. Existing validation studies are predominantly based on synthetic datasets (7); this raises concerns regarding the generalizability and clinical reliability of their results.

Given this gap, our study aimed to evaluate the efficacy of TabPFN in classifying benign and malignant renal tumors based on radiomic features extracted from computed tomography (CT) images. The study also aimed to compare TabPFN with established ML algorithms using two independent datasets, to assess its real-world clinical performance and feasibility as a radiomics-based classification tool. We present this article in accordance with the CLEAR reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1132/rc).

Methods

Ethical approval

The study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. This study was approved by the Medical Ethics Committee of the Guangdong Provincial Hospital of Chinese Medicine (approval Nos. ZE2023-090-01 for dataset A and ZE2024-294-01 for dataset B). The Ethics Committee of the Provincial Hospital of Chinese Medicine waived the requirement for written informed consent because the study involved a retrospective analysis of de-identified imaging and clinical data, posed no more than minimal risk to participants, and it was impracticable to obtain consent from all individuals due to the time span of data collection. All datasets were anonymised before analysis and no additional procedures were performed (9,10). All participating institutions were also informed and agreed the study.

Collection of data

This retrospective study comprised two datasets. The first (dataset A) included data from 207 cystic renal masses (CRMs; 92 benign and 115 malignant) classified under the Bosniak II F category (9,11). The second dataset (dataset B) included 92 cases, of which 41 were renal oncocytomas (ROs) and 51 were chromophobe renal cell carcinomas (CRCCs), respectively (10).

Dataset A was sourced from two centers of the Guangdong Provincial Hospital of Chinese Medicine, namely, center 1 in Guangzhou and center 2 in Zhuhai; data collection was performed between January 2018 and February 2022. The

scanning parameters for this dataset have been detailed in a previous study (9). Cases that met the following criteria were considered eligible for inclusion: (I) both unenhanced and contrast-enhanced CT images, including those from the corticomedullary phase (CMP) and nephrographic phase (NP), were available for classifying CRMs according to the Bosniak system; (II) had comprehensive clinical information including patient age, sex, lesion location, complete surgical and/or biopsy records, and histopathological results retrieved from the pathology databases of both centers; and (III) high-quality CT images obtained from the picture archiving and communication system. Cases with poor-quality or incomplete CT datasets and those with lesions classified as Bosniak category II or lower were excluded (9). During dataset allocation, cases from center 1 (77 benign and 85 malignant CRMs) were designated as the training set, whereas those from center 2 (15 benign and 30 malignant CRMs) were utilized for external validation (9).

Dataset B was obtained from five medical centers in Guangdong Province, namely, Guangdong Hospital of Traditional Chinese Medicine, Guangzhou (center 1); Guangdong Hospital of Traditional Chinese Medicine, Zhuhai (center 2); First Affiliated Hospital of Sun Yat-sen University, Guangzhou (center 3); Affiliated Panyu Central Hospital of Guangdong Medical University, Guangzhou (center 4); and Longgang Central Hospital, Shenzhen (center 5) (10). The data were collected between January 2018 and July 2024 based on scanning parameters mentioned in our previous study (10). The following cases were considered eligible for inclusion in this dataset: (I) where enhanced CT scans, and specifically images from the CMP and NP phases, were available; (II) where complete clinical records, including patient demographic data (such as age and sex), lesion location, surgical details (radical or partial nephrectomy), and histopathological and immunohistochemical findings (obtained at each center) were available; and (III) where high-quality CT images were available from the picture archiving and communication system. Cases with incomplete or low-quality CT images and those where the CT scans did not fully capture the lesion were excluded (10). After randomization, the cases were divided into training (36 and 29 cases of CRCC and RO, respectively) and validation (15 and 12 cases of CRCC and RO, respectively) sets in a 7:3 ratio (10).

Image segmentation and radiomic feature acquisition

For dataset A, mass segmentation and radiomic feature extraction were conducted using the 3D Slicer open-source platform (version 5.2.1; https://www.slicer.org/). A total of 855 radiomic features were extracted using the PyRadiomics package, integrated within 3D Slicer (11,12). These features were organized into seven categories: first-order statistics, two-dimensional features, gray-level co-occurrence matrix (GLCM), gray-level dependence matrix (GLDM), gray-level size-zone matrix (GLSZM), gray-level run-length matrix (GLRLM), and neighboring gray-tone difference matrix (NGTDM) (9,11). Additionally, 14 filters were applied to the original images, including exponential, gradient, square, square root, logarithm, and Ibp2D filters, and various wavelet transformations (HLH, HLL, LHL, LLL, LHH, LLH, HHL, and HHH) (9,11).

For dataset B, image segmentation was performed using ITK-SNAP (http://www.itksnap.org), another open-source tool (13). To ensure standardized image resolution, all images and segmentation masks were resampled to a voxel size of 1×1×1 mm^3 (10). The PyRadiomics package in Python was then employed to extract 2,260 radiomic features from the CMP and NP images; these included shape, texture, first-order statistics, Laplacian of Gaussian, GLCM, GLRLM, GLSZM, NGTDM, and GLDM features (10). The same 14 filters used in dataset A were applied, and feature normalization was conducted using z-score standardization (10).

In both datasets, tumor segmentation was exclusively performed manually by two highly experienced associate chief radiologists (T.L. and L.H.), each having over 15 years of specialized experience in abdominal imaging. Accurate and complete tumor boundary delineation was ensured throughout the process, and special care was taken to avoid excessive margins or incomplete segmentation (9,10).

We performed detailed intra- and inter-observer reliability analyses to rigorously evaluate the reproducibility of extracted radiomic features. In particular, we randomly selected 80 cases from dataset A and 40 cases from dataset B for repeated manual segmentation by two additional experienced radiologists (S.X. and G.Z.). We then quantitatively assessed reproducibility using the intraclass correlation coefficient (ICC). Only radiomic features with ICC values greater than 0.75 (indicating excellent reproducibility) were included in the final radiomic analyses (9-11).

After meeting the standard of consistency, the features were further selected to avoid overfitting. The least absolute shrinkage and selection operator (LASSO) method was applied to select the most suitable radiomic features to develop a radiomic signature using the scikit-learn package

in Python. First, tenfold cross-validation was performed over 1,000 iterations to obtain the optimal regularization parameter (λ). Second, the selected LASSO model was used to calculate the coefficient of each feature, and features with non-zero coefficients were retained. Finally, a sequential feature selection method with both forward and backward steps was applied using the mlxtend package to further refine the LASSO-selected features.

ML algorithms

To ensure objective comparison, seven ML algorithms were used to develop the radiomic models. In addition to the transformer-based TabPFN algorithm, six conventional models were used; these included SVM, stochastic gradient descent (SGD), k-nearest neighbor (KNN), RF, extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM) algorithms. Standardized Python packages (scikit-learn: SVM, SGD, KNN, and RF; XGBoost; and LightGBM) and the TabPFN GitHub repository (https://github.com/PriorLabs/TabPFN) were used.

The mentioned algorithms were applied to develop radiomic models of unenhanced, CMP, and NP images from dataset A (11), and CMP, NP, and fusion images (combining CMP and NP) from dataset B (10). For the traditional algorithms, hyperparameter optimization was performed rigorously via grid search and 10-fold cross-validation; manual tuning could be avoided for TabPFN owing to its pre-trained transformer architecture (pre-trained on 1.2 million synthetic datasets simulating radiomic feature distributions). This unique feature of TabPFN helped in dynamically adjusting attention weights during inference through 12 transformer layers (eight attention heads), thereby directly modeling global dependencies between multi-phase CT features. The workflow of the radiomic approach is illustrated in *Figure 1*.

Statistical analysis

Statistical analyses were performed using SPSS (version 26.0, IBM, Armonk, NY, USA) and Python (version 3.7.1) software. Python was used for feature extraction, selection, model development, and validation, whereas SPSS was used for cohort comparisons. All statistical tests were two-sided, and p-values below 0.05 were considered statistically significant (9-11).

The normality of clinical data distribution was assessed using the Shapiro-Wilk test. Continuous variables with normal distribution have been presented as the mean ± standard deviation (SD), while non-normally distributed variables have been presented as the median and interquartile range (IQR); categorical variables have been reported as counts and percentages. For group comparisons, the χ^2 test of independence was used for categorical variables, and independent-sample *t*- or Wilcoxon tests were employed for continuous variables (9-11). The ICC was calculated to evaluate the reproducibility of the radiomic features extracted by different radiologists (10,11). In addition, the discriminative ability of the models was assessed using receiver operating characteristic (ROC) analysis; the area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were used as key performance metrics (10,11). AUC values across different radiomic models were compared using the DeLong test, and the sklearn.calibration package in Python was employed along with custom scripts to generate calibration curves, conduct decision curve analysis (DCA) (9,10), and visualize the sensitivity, specificity, accuracy, PPV, and NPV derived from ROC analysis. Feature importance in the ML models was analyzed and visualized using the sklearn.inspection package (10,11).

Results

Patient characteristics

Dataset A comprised 207 patients (105 males and 102 females; mean age: 59.1±11.5 years) diagnosed with CRMs. Among them, 92 (51 males and 41 females; mean age: 58.0±13.7 years) had benign CRMs, whereas 115 (54 males and 61 females; mean age: 59.8±11.4 years) had malignant CRMs (11). Dataset B included 92 patients (38 males and 54 females; mean age: 53.6±13.65 years) diagnosed with RO and CRCC. This subset consisted of 41 patients (16 males and 25 females; mean age: 52.4±11.44 years) with RO and 51 patients (22 males and 29 females; mean age: 54.0±14.8 years) with CRCC (10).

The analysis revealed no significant differences between patients with benign and malignant lesions from either dataset in terms of age, sex, and lesion location and size (10,11). The patient characteristics are summarized in Tables S1,S2.

Radiomic feature selection

Univariate analyses, LASSO selection, 10-fold cross-

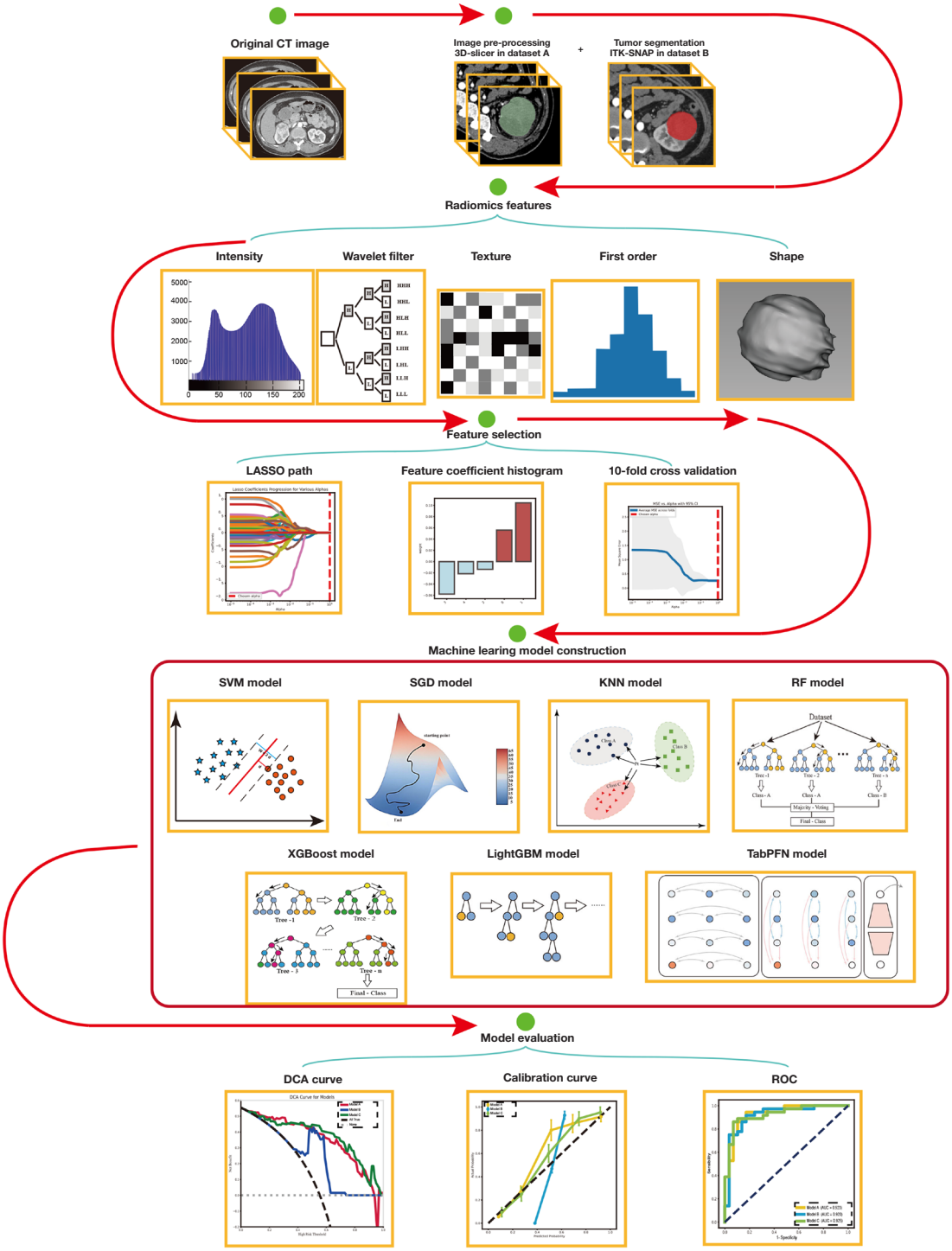

Figure 1 Workflow for development and evaluation of the ML classifiers. AUC, area under the curve; CI, confidence interval; CT, computed tomography; DCA, decision curve analysis; KNN, k-nearest neighbor; LASSO, least absolute shrinkage and selection operator; LightGBM, light gradient boosting machine; ML, machine learning; MSE, mean squared error; RF, random forest; ROC, receiver operating characteristic; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

validation, bidirectional elimination, and ICCs were performed for feature selection. In dataset A, this process yielded 4, 2, and 5 key features from unenhanced, CMP, and NP CT images, respectively; in dataset B, 9, 3, and 12 features were extracted from CMP, NP, and fusion (CMP + NP) CT images, respectively (9-11). The identified radiomic features are summarized in Table S3.

Diagnostic performance of the ML algorithms

In dataset A, all ML algorithms demonstrated excellent performance in the training sets of unenhanced (AUC: 0.942–1.000), CMP (AUC: 0.989–1.000), and NP (AUC: 0.935–1.000) models. In addition, most ML algorithms performed satisfactorily in the validation sets of unenhanced (AUC: 0.821–0.902), CMP (AUC: 0.875–0.922), and NP (AUC: 0.800–0.946) models. However, the performance of SGD (AUC: 0.800) was slightly poorer than that of the other ML algorithms in the NP model. The TabPFN algorithm achieved mid-to-high ranking performance in the validation sets of unenhanced (AUC: 0.875) and CMP (AUC: 0.900) models and demonstrated the best performance in the NP model (AUC: 0.946).

In dataset B, RF (training and validation AUC: 0.964 and 0.794, respectively), XGBoost (training and validation AUC: 1.000 and 0.694, respectively), LightGBM (training and validation AUC: 0.997 and 0.683, respectively), and TabPFN (training and validation AUC: 0.958 and 0.800, respectively) algorithms performed better than the SVM (training and validation AUC: 0.779 and 0.606, respectively) and KNN (training and validation AUC: 0.778 and 0.639, respectively) algorithms in both training and validation sets of CMP models. However, the SGD (both training and validation AUC: 0.500) algorithm failed to achieve convergence. All the ML algorithms demonstrated similar and satisfactory performance in both training and validation sets of the NP model (training AUC: 0.803–0.915, validation AUC: 0.581–0.772); that of TabPFN (training AUC: 0.841, validation AUC: 0.700) was in the mid-range. Except for SGD, all other ML algorithms demonstrated excellent performance in the training sets of the fusion model; the performance of TabPFN (AUC: 0.783) was in the upper range and slightly lower than that of RF (AUC: 0.811) in the validation set. However, SGD (both training and validation AUC: 0.500) persistently failed to achieve convergence. The diagnostic performance of the algorithms is summarized in *Tables 1*,*2*, and the results of the DeLong test are presented in Tables S4,S5.

Calibration curve analysis and DCA were conducted for the ML classifiers across both training and validation sets of datasets A and B. In dataset A, the calibration curves were closely aligned with the ideal line in the training set. However, there were varying degrees of deviations among ML classifiers in the validation set. DCA for the validation set indicated comparable performance across all ML classifiers, except for SGD in the NP model, which underperformed relative to the others.

In dataset B, SGD failed to converge in both the training and validation sets for the CMP and fusion models. The calibration curves for the other ML classifiers remained relatively close to the ideal line in the training sets. However, during validation, significant deviations were noted for SVM (in the CMP and fusion models) and LightGBM (in the NP model). DCA for the validation set revealed similar performance among all ML classifiers except for SGD (in the CMP and fusion models). The ROC, calibration, and DCA curves are illustrated in *Figures 2*,*3*.

Feature importance analysis of the ML models

The feature importance plots of the ML models are shown in Figures S1,S2. In general, SVM, SGD, and KNN tended to assign weights to a few features; conversely, RF, XGBoost, and LightGBM assigned weights to more features in datasets A and B. Although TabPFN demonstrated greater similarity to RF, XGBoost, and LightGBM algorithms in more cases, it reduced the weight of insignificant features (particularly in the CMP and fusion models of dataset B). The results of the feature importance analysis are shown in Figure S1 (for CRM) and Figure S2 (for CRCC and RO).

Discussion

In this study, we used two independent datasets to evaluate the diagnostic efficiency of the TabPFN algorithm and compare it with that of other established ML algorithms. To the best of our knowledge, this study is the first to apply the TabPFN algorithm in the context of renal tumor radiomics and evaluate its performance using real-world clinical datasets. This represents an important step toward the translation of transformer-based tabular deep learning models into clinical practice (14). TabPFN not only exhibited commendable performance in the dataset for CRMs (dataset A) but also demonstrated its unique capability in dynamic feature weight allocation. Unlike traditional algorithms that either overemphasized dominant

Table 1 Sensitivity, specificity, accuracy, and AUC of radiomics models in the training and validation sets of dataset A

ML algorithm	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Training set for U						
SVM	89.47 (80.58–94.57)	98.80 (93.49–99.79)	94.34 (89.59–96.99)	98.55 (92.24–99.74)	91.11 (83.43–95.43)	0.981 (0.957–1.00)
SGD	75.00 (64.22–83.37)	96.47 (90.13–98.79)	86.34 (80.18–90.80)	95.00 (86.30–98.29)	81.19 (72.48–87.61)	0.942 (0.891–0.993)
KNN	100 (95.19–100)	100 (95.68–100)	100 (97.67–100)	100 (95.19–100)	100 (95.68–100)	1.00 (1.00–1.00)
RF	93.42 (85.51–97.16)	96.47 (90.13–98.79)	95.03 (90.50–97.46)	95.95 (88.75–98.61)	94.25 (87.24–97.52)	0.990 (0.979–1.00)
XGBoost	92.11 (83.83–96.33)	98.82 (93.63–99.79)	95.65 (91.30–97.88)	98.59 (92.44–99.75)	93.33 (86.21–96.91)	0.997 (0.994–1.00)
LightGBM	93.42 (85.51–97.16)	97.65 (91.82–99.35)	95.65 (91.30–97.88)	97.26 (90.55–99.25)	94.32 (87.38–97.55)	0.990 (0.979–1.00)
TabPFN	93.42 (85.51–97.16)	97.65 (91.82–99.35)	95.65 (91.30–97.88)	97.26 (90.55–99.25)	94.32 (87.38–97.55)	0.993 (0.984–1.00)
Validation set for U						
SVM	71.43 (50.04–86.19)	83.33 (66.44–92.66)	78.43 (65.37–87.51)	75.00 (53.13–88.81)	80.65 (63.72–90.81)	0.894 (0.805–0.981)
SGD	71.43 (50.04–86.19)	90.00 (74.38–96.54)	82.35 (69.75–90.43)	83.33 (60.78–94.16)	81.82 (65.61–91.39)	0.821 (0.690–0.951)
KNN	80.95 (60.00–92.33)	80.00 (62.69–90.50)	80.39 (67.54–88.98)	73.91 (53.53–87.45)	85.71 (68.51–94.30)	0.873 (0.784–0.962)
RF	85.71 (65.36–95.02)	73.33 (55.55–85.82)	78.43 (65.37–87.51)	69.23 (50.01–83.50)	88.00 (70.04–95.83)	0.902 (0.815–0.989)
XGBoost	80.95 (60.00–92.33)	80.00 (62.69–90.50)	80.39 (67.54–88.98)	73.91 (53.53–87.45)	85.71 (68.51–94.30)	0.873 (0.778–0.968)
LightGBM	76.19 (54.91–89.37)	80.00 (62.69–90.50)	78.43 (65.37–87.51)	72.73 (51.85–86.85)	82.76 (65.45–92.40)	0.863 (0.771–0.955)
TabPFN	66.67 (45.37–82.81)	93.33 (78.68–98.15)	82.35 (69.75–90.43)	87.50 (63.98–96.50)	80.00 (64.11–89.96)	0.875 (0.772–0.978)
Training set for CMP						
SVM	98.70 (93.00–99.77)	92.94 (85.44–96.72)	95.68 (91.35– 97.89)	92.68 (84.94–96.60)	98.75 (93.25–99.78)	0.993 (0.985–1.00)
SGD	92.21 (84.02–96.38)	97.65 (91.82–99.35)	95.06 (90.56–97.48)	97.26 (90.55–99.25)	93.26 (86.06–96.87)	0.989 (0.976–1.00)
KNN	100 (95.25–100)	100 (95.68–100)	100 (97.68–100)	100 (95.25–100)	100 (95.68–100)	1.00 (1.00–1.00)
RF	97.40 (91.02–99.28)	96.47 (90.13–98.79)	96.91 (92.98–98.67)	96.15 (89.29–98.68)	97.62 (91.73–99.34)	0.992 (0.984–1.00)
XGBoost	97.40 (91.02–99.28)	96.47 (90.13–98.79)	96.91 (92.98–98.67)	96.15 (89.29–98.68)	97.62 (91.73–99.34)	0.997 (0.994–1.00)
LightGBM	97.40 (91.02–99.28)	96.47 (90.13–98.79)	96.91 (92.98–98.67)	96.15 (89.29–98.68)	97.62 (91.73–99.34)	0.995 (0.991–1.00)
TabPFN	96.10 (89.16–98.67)	96.47 (90.13–98.79)	96.30 (92.16–98.29)	96.10 (89.16–98.67)	96.47 (90.13–98.79)	0.993 (0.985–1.00)

Table 1 (*continued*)

Table 1 (*continued*)

ML algorithm	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Validation set for CMP						
SVM	95.24 (77.33–99.15)	70.00 (52.12–83.34)	80.39 (67.56–88.98)	68.97 (50.77–82.72)	95.45 (78.20–99.19)	0.910 (0.823–0.996)
SGD	80.95 (60.00–92.33)	86.67 (70.32–94.69)	84.31 (71.99–91.83)	80.95 (60.00–92.33)	86.67 (70.32–94.69)	0.922 (0.848–0.993)
KNN	85.71 (65.36–95.02)	73.33 (55.55–85.82)	78.43 (65.37–87.51)	69.23 (50.01–83.50)	88.00 (70.04–95.83)	0.875 (0.791–0.957)
RF	95.24 (77.33–99.15)	73.33 (55.55–85.82)	82.35 (69.75–90.43)	71.43 (52.94–84.75)	95.65 (79.01–99.23)	0.910 (0.827–0.989)
XGBoost	95.24 (77.33–99.15)	73.33 (55.55–85.82)	82.35 (69.75–90.43)	71.43 (52.94–84.75)	95.65 (79.01–99.23)	0.896 (0.824–0.981)
LightGBM	95.24 (77.33–99.15)	73.33 (55.55–85.82)	82.35 (69.75–90.43)	71.43 (52.94–84.75)	95.65 (79.01–99.23)	0.903 (0.805–0.974)
TabPFN	100 (84.54–100)	70.00 (52.12–83.34)	82.35 (69.75–90.43)	70.00 (52.12–83.34)	100 (100–100)	0.900 (0.807–0.993)
Training set for NP						
SVM	98.70 (93.00–99.77)	97.65 (91.82–99.35)	98.15 (94.70–99.37)	97.44 (91.12–99.29)	98.81 (93.56–99.79)	0.997 (0.993–1.00)
SGD	98.70 (93.00–99.77)	88.24 (79.68–93.48)	93.21 (88.25–99.77)	88.37 (79.90–93.56)	98.68 (92.92–99.77)	0.935 (0.911–0.959)
KNN	97.40 (91.02–99.28)	97.65 (91.82–99.35)	97.53 (93.82–99.04)	97.40 (91.02–99.28)	97.65 (91.82–99.35)	0.998 (0.994–1.00)
RF	98.70 (93.00–99.77)	97.65 (91.82–99.35)	98.15 (94.70–99.37)	97.44 (91.12–99.29)	98.81 (93.56–99.79)	0.999 (0.998–1.00)
XGBoost	97.40 (91.02–99.28)	98.82 (93.63–99.79)	98.15 (94.70–99.37)	98.68 (96.92–99.77)	97.67 (91.91–99.36)	0.994 (0.981–1.00)
LightGBM	100 (95.25–100)	100 (95.68–100)	100 (97.68–100)	100 (95.25–100)	100 (95.68–100)	1.00 (1.00–1.00)
TabPFN	100 (95.25–100)	98.82 (93.63–99.79)	99.38 (96.59–99.89)	98.72 (93.09–99.77)	100 (95.63–100)	0.999 (0.998–1.00)
Validation set for NP						
SVM	95.24 (77.33–99.15)	73.33 (55.55–85.82)	82.35 (69.75–90.43)	71.43 (52.94–84.75)	95.65 (79.01–99.23)	0.933 (0.866–1.00)
SGD	100 (84.54–100)	60.00 (42.32–75.41)	76.47 (63.24–86.00)	63.64 (46.62–77.81)	100 (82.41–100)	0.800 (0.712–0.888)
KNN	95.24 (77.33–99.15)	76.67 (59.07–88.21)	84.31 (71.99–91.83)	74.07 (55.32–86.83)	95.83 (79.76–99.26)	0.918 (0.845–0.991)
RF	90.48 (71.09–97.35)	76.67 (59.07–88.21)	82.35 (69.75–90.43)	73.08 (53.92–86.30)	92.00 (75.03–97.78)	0.939 (0.876–1.00)
XGBoost	90.48 (71.09–97.35)	80.00 (62.69–90.50)	84.31 (71.99–91.83)	76.00 (56.57–88.50)	92.31 (75.86–97.86)	0.886 (0.804–0.999)
LightGBM	95.24 (77.33–99.15)	80.00 (62.69–90.50)	86.27 (74.28–83.19)	76.92 (59.95–88.97)	96.00 (80.46–99.29)	0.937 (0.875–1.00)
TabPFN	95.24 (77.33–99.15)	80.00 (62.69–90.50)	86.27 (74.28–83.19)	76.92 (59.95–88.97)	96.00 (80.46–99.29)	0.946 (0.886–1.00)

Data are presented as value (95% CI). AUC, area under the curve; CI, confidence interval; CMP, corticomedullary phase; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; NP, nephrographic phase; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; U, unenhanced; XGBoost, extreme gradient boosting.

Table 2 Sensitivity, specificity, accuracy, and AUC of radiomics models in the training and validation sets of dataset B

ML algorithm	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Training set for CMP						
SVM	91.67 (78.17–97.13)	41.38 (25.51–59.26)	69.23 (57.20–79.11)	66.00 (52.15–77.56)	80.00 (54.81–92.95)	0.779 (0.664–0.893)
SGD	100 (100–100)	0 (0–0)	55.38 (43.30–67.46)	55.38 (43.30–67.46)	N/A	0.500 (0.500–0.500)
KNN	72.22 (56.01–84.15)	65.52 (47.34–80.06)	69.23 (57.20–79.11)	72.22 (56.01–84.15)	65.52 (47.34–80.06)	0.778 (0.664–0.893)
RF	94.44 (81.86–98.46)	86.21 (69.44–94.50)	90.77 (81.29–95.70)	89.47 (75.87–95.83)	92.59 (76.63–97.94)	0.964 (0.919–1.00)
XGBoost	100 (90.36–100)	100 (88.30–100)	100 (94.42–100)	100 (90.36–100)	100 (88.30–100)	1.00 (1.00–1.00)
LightGBM	100 (90.36–100)	96.55 (82.82–99.39)	98.46 (91.79–99.73)	97.30 (86.18–99.52)	100 (87.94–100)	0.997 (0.990–1.00)
TabPFN	94.44 (81.86–98.46)	86.21 (69.44–94.50)	90.77 (81.29–95.70)	89.47 (75.87–95.83)	92.59 (76.63–97.94)	0.958 (0.911–1.00)
Validation set for CMP						
SVM	100 (79.61–100)	16.67 (4.70–44.80)	62.96 (44.23–78.47)	60.00 (40.74–76.60)	100 (34.24–100)	0.606 (0.343–0.868)
SGD	100 (100–100)	0 (0–0)	55.56 (36.82–74.30)	55.56 (36.82–74.30)	N/A	0.500 (0.500–0.500)
KNN	73.33 (48.05–89.10)	50.00 (25.38–74.62)	62.96 (44.23–78.47)	64.71 (41.30–82.69)	60.00 (31.27–83.18)	0.639 (0.407–0.871)
RF	80.00 (54.81–92.95)	58.33 (31.95–80.67)	70.37 (51.52–84.15)	70.59 (46.87–86.72)	70.00 (39.68–89.22)	0.794 (0.625–0.969)
XGBoost	86.67 (62.12–96.26)	41.67 (19.33–68.05)	66.67 (47.82–81.36)	65.00 (43.29–81.88)	71.43 (35.89–91.78)	0.694 (0.484–0.905)
LightGBM	93.33 (83.92–100)	41.67 (23.07–60.27)	70.37 (53.15–87.59)	70.37 (53.15–87.59)	83.33 (69.27–97.39)	0.683 (0.455–0.912)
TabPFN	86.67 (62.12–96.26)	50.00 (25.38–74.62)	70.37 (51.52–84.15)	68.42 (46.01–84.64)	75.00 (40.93–92.85)	0.800 (0.632–0.968)
Training set for NP						
SVM	75.00 (58.93–86.25)	79.31 (61.61–90.15)	76.92 (65.36–85.49)	81.82 (65.61–91.39)	71.88 (54.63–84.44)	0.801 (0.688–0.914)
SGD	72.22 (56.01–84.15)	79.31 (61.61–90.15)	75.38 (63.69–84.24)	81.25 (64.69–91.11)	69.70 (52.66–82.62)	0.803 (0.691–0.915)
KNN	75.00 (58.93–86.25)	82.76 (65.45–92.40)	78.46 (67.03–86.71)	84.38 (68.25–93.14)	72.73 (55.78–84.93)	0.821 (0.720–0.923)
RF	86.11 (71.34–93.92)	72.41 (54.28–85.30)	80.00 (68.73–87.92)	79.49 (64.47–89.22)	80.77 (62.12–91.49)	0.874 (0.775–0.978)
XGBoost	94.44 (81.86–98.46)	68.97 (50.77–82.72)	83.08 (72.18–90.28)	79.07 (64.79 –88.58)	90.91 (72.18–97.47)	0.912 (0.839–0.996)
LightGBM	80.56 (64.97–90.25)	82.76 (65.45–92.40)	81.54 (70.45–89.11)	85.29 (69.87–93.55)	77.42 (60.19–88.61)	0.915 (0.838–1.00)
TabPFN	77.78 (61.91–88.28)	79.31 (61.61–90.15)	78.46 (67.03–86.71)	82.35 (66.49–91.65)	74.19 (56.75–86.30)	0.841 (0.741–0.940)

Table 2 (*continued*)

Table 2 (*continued*)

ML algorithm	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)	AUC
Validation set for NP						
SVM	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.761 (0.561–0.962)
SGD	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.750 (0.541–0.960)
KNN	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.772 (0.597–0.947)
RF	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.678 (0.445–0.908)
XGBoost	53.33 (30.12–75.19)	66.67 (39.06–86.19)	59.26 (40.73–75.49)	66.67 (39.06–86.19)	66.67 (39.06–86.19)	0.606 (0.381–0.830)
LightGBM	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.581 (0.352–0.804)
TabPFN	46.67 (24.81–69.88)	83.33 (55.20–95.30)	62.96 (44.23–78.47)	77.78 (45.26–93.68)	55.56 (33.72–75.44)	0.700 (0.494–0.906)
Training set for fusion						
SVM	77.78 (61.91–88.28)	79.31 (61.61–90.15)	78.46 (67.03–86.71)	82.35 (66.49–91.65)	74.19 (56.75–86.30)	0.882 (0.795–0.969)
SGD	0 (0–0)	100 (100–100)	44.62 (32.54–56.70)	N/A	44.62 (32.54–56.70)	0.500 (0.500–0.500)
KNN	72.22 (56.01–84.15)	72.41 (54.28–85.30)	72.31 (60.42–81.71)	76.47 (60.00–87.56)	67.74 (50.14–81.43)	0.878 (0.812–0.943)
RF	94.44 (81.86–98.46)	89.66 (73.61–96.42)	92.31 (83.22–96.67)	91.89 (78.70–97.20)	92.86 (77.35–98.02)	0.951 (0.922–1.00)
XGBoost	100 (90.36–100)	100 (88.30–100)	100 (94.42–100)	100 (90.36–100)	100 (88.30–100)	1.00 (1.00–1.00)
LightGBM	97.22 (85.83–99.51)	96.55 (82.82–99.39)	96.92 (89.46–99.15)	97.22 (85.83–99.51)	96.55 (82.82–99.39)	0.998 (0.996–1.00)
TabPFN	100 (90.36–100)	100 (88.30–100)	100 (94.42–100)	100 (90.36–100)	100 (88.30–100)	1.00 (1.00–1.00)
Validation set for fusion						
SVM	80.00 (54.81–92.95)	66.67 (39.06–86.19)	74.07 (55.32–86.83)	75.00 (50.50–89.82)	72.73 (43.43–90.25)	0.689 (0.465–0.913)
SGD	0 (0–0)	100 (100–100)	44.44 (25.69–63.18)	N/A	44.44 (25.69–63.18)	0.500 (0.500–0.500)
KNN	66.67 (41.71–84.82)	50.00 (25.38–74.62)	59.26 (40.73–75.49)	62.50 (38.64–81.52)	54.55 (28.01–78.73)	0.697 (0.482–0.918)
RF	80.00 (54.81–92.95)	58.33 (31.95–80.67)	70.37 (51.52–84.15)	70.59 (46.87–86.72)	70.00 (39.68–89.22)	0.811 (0.630–0.997)
XGBoost	80.00 (54.81–92.95)	50.00 (25.38–74.62)	66.67 (47.82–81.36)	66.67 (43.75–83.72)	66.67 (35.42–87.94)	0.744 (0.553–0.936)
LightGBM	60.00 (35.75–80.18)	58.33 (31.95–80.67)	59.26 (40.73–75.49)	64.29 (38.76–83.66)	53.85 (29.14–76.79)	0.672 (0.456–0.890)
TabPFN	80.00 (54.81–92.95)	66.67 (39.06–86.19)	74.07 (55.32–86.83)	75.00 (50.50–89.82)	72.73 (43.43–90.25)	0.783 (0.602–0.965)

Data are presented as value (95% CI). AUC, area under the curve; CI, confidence interval; CMP, corticomedullary phase; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; N/A, not available; NP, nephrographic phase; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

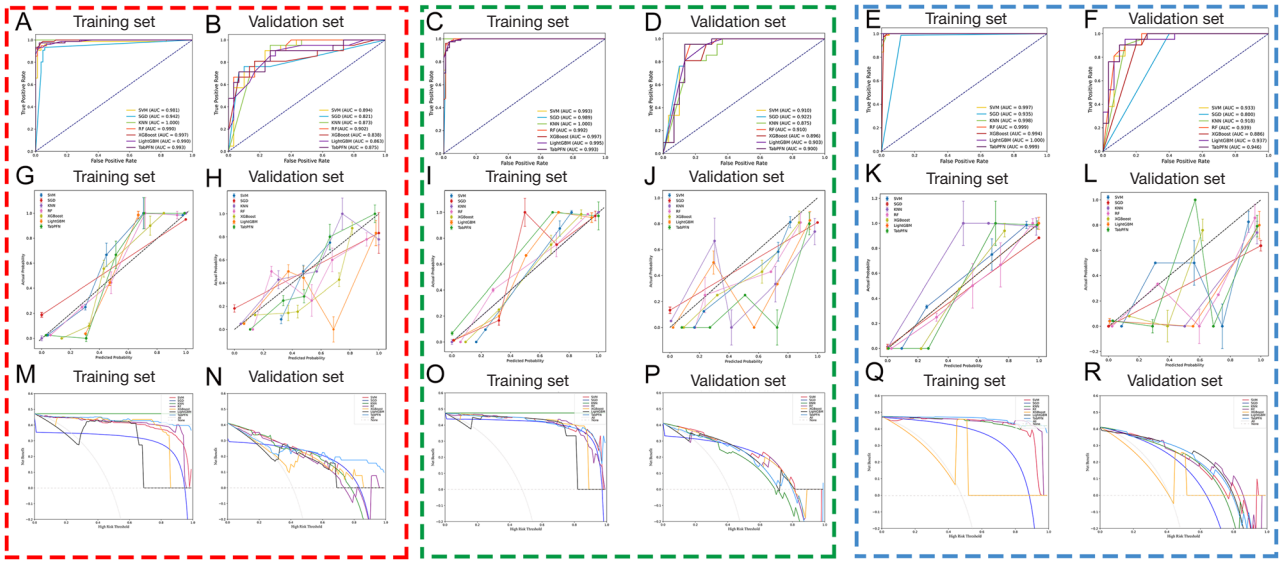

Figure 2 Evaluation of the ML classifiers based on ROC analysis [(A,B) unenhanced; (C,D) CMP; (E,F) NP], calibration curves [(G,H) unenhanced; (I,J) CMP; (K,L) NP], and DCA [(M,N) unenhanced; (O,P) CMP; (Q,R) NP] of the training and validation sets of dataset A. AUC, area under the curve; CMP, corticomedullary phase; DCA, decision curve analysis; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; NP, nephrographic phase; RF, random forest; ROC, receiver operating characteristic; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

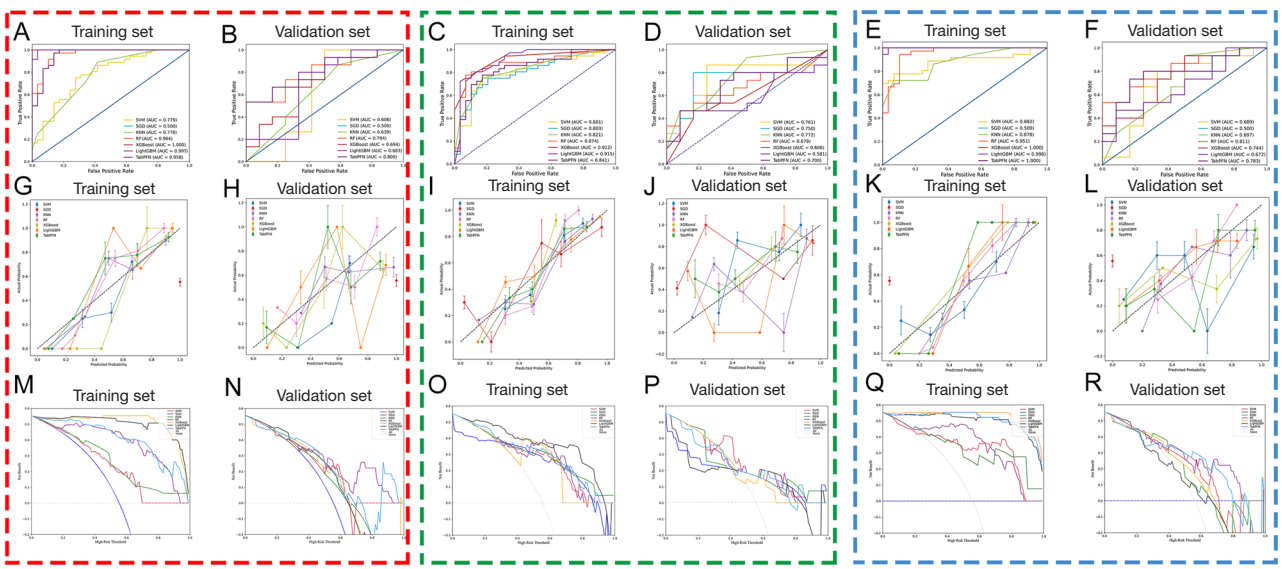

Figure 3 Evaluation of the ML classifiers based on ROC analysis [(A,B) unenhanced; (C,D) CMP; (E,F) NP], calibration curves [(G,H) unenhanced; (I,J) CMP; (K,L) NP], and DCA [(M,N) unenhanced; (O,P) CMP; (Q,R) NP] of the training and validation sets of dataset B. AUC, area under the curve; CMP, corticomedullary phase; DCA, decision curve analysis; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; NP, nephrographic phase; RF, random forest; ROC, receiver operating characteristic; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

features (e.g., SVM/KNN) (15,16) or required explicit regularization (e.g., SGD) (17,18), TabPFN inherently prioritized radiomic features with higher discriminative power (while adaptively suppressing noisy or redundant ones) (14); this is attributable to its transformer-based architecture. This intelligent weight distribution, achieved without manual hyperparameter tuning, contributed to its robust performance across the unenhanced CMP and NP datasets (mid- to upper-range AUCs). Importantly, TabPFN circumvented the convergence failures observed in SGD and mitigated the overfitting risks of ensemble methods in dataset B; notably, the limited sample size (n=92) and relatively high feature dimensionality (9 and 12 radiomic features) in this dataset created challenges. Owing to the benefits offered by the pre-trained attention mechanisms, TabPFN maintained stable decision boundaries even when confronting imbalanced feature importance distributions; this represents a critical advantage for clinical applications where data are sparse and heterogeneous.

The TabPFN algorithm, developed by Hollmann *et al.* (7) in January 2025, represents a novel ML model optimized for tabular data. It significantly reduces training time while enabling fine-tuning, data generation, density estimation, and learning reusable embeddings. The core innovation lies in generating a large synthetic tabular dataset corpus, which trains a transformer-based neural network to solve diverse prediction tasks. Unlike conventional methods, TabPFN employs cross-dataset training and applies inference to entire datasets rather than individual samples. Prior to real-world deployment, the model underwent pretraining on millions of synthetic datasets that represented varied prediction tasks. However, as mentioned in a recent article, its application to real-world datasets remains limited (7). This raises concerns regarding its ability to fully anticipate clinical research scenarios and indicates a need for rigorous real-world validation. It also highlights the need for comparative studies with existing algorithms for future clinical adoption.

Based on these considerations, we evaluated the performance of the TabPFN algorithm in our previously obtained datasets. In this context, datasets A and B represent distinct research scenarios in this radiomics algorithm study. The former included 213 patients with CRMs, and radiomic features were directly extracted using 3D Slicer software. The sample size aligned with standard radiomic requirements, and each scanning phase contained a moderate number of features. Dataset B included 91 patients with CRCC and RO, where radiomic features were derived using ITK-SNAP and Python-based analysis. As it included rare pathological subtypes, this dataset presented specific challenges. It also included smaller sample sizes for each subtype and higher-dimensional radiomic features; this further tested algorithm convergence capabilities and gradient descent performance (13).

The TabPFN algorithm demonstrated superior performance in dataset A. It consistently ranked in the upper-middle tier across all subsets, with particularly notable results in the NP model. The algorithm exhibited stable and satisfactory outcomes in dataset B and successfully navigated challenges posed by limited sample sizes and high feature dimensionality. Notably, TabPFN did not exhibit the convergence failures observed with SGD (19), and it outperformed algorithms such as SVM and KNN (20), which showed poor convergence. Overall, the performance of TabPFN was at par with that of traditional methods such as RF, LightGBM, and XGBoost.

Although TabPFN demonstrated exceptional performance in dataset A, its stability in dataset B was particularly noteworthy. The findings showed that TabPFN achieved consistent convergence and maintained stable performance, even under the challenging conditions of small sample sizes (n=91) and high-dimensional radiomic features. This contrasts sharply with the performance of traditional algorithms such as SGD (which exhibited non-convergence) and SVM/KNN (which showed poor convergence) under similar constraints (20). The model showed good reliability in dataset B, which included rare clinical cases (CRCC and RO) and a complex feature space; this aligned with its design principles as a tabular foundation model, which is pre-trained on millions of synthetic datasets. Notably, the transformer-based architecture and bidirectional attention mechanisms of the TabPFN algorithm enable efficient handling of feature-sample interactions and mitigate common pitfalls of high-dimensional low-sample scenarios (7,14). This shows its reliability in real-world clinical research, particularly for limited cohort sizes with rich radiomic feature sets. The findings also highlight its robustness in handling both conventional and high-dimensional radiomics datasets, consistent with its design principles for small-to-medium scale tabular data (≤10,000 samples).

The findings from feature importance analysis revealed the stability of the TabPFN algorithm across both datasets A and B. Traditional models such as the SVM, SGD, and KNN typically focus feature weights on one or two dominant features. Although this approach can yield satisfactory diagnostic performance in moderately sized

datasets (such as dataset A) (21,22), it often results in suboptimal convergence in scenarios characterized by high dimensionality and low sample sizes (such as dataset B).

In contrast, ensemble methods such as RF (23), XGBoost (24), and LightGBM (25) tend to distribute weights more evenly across a broader range of features. Although TabPFN aligns with these ensemble methods, it also demonstrates a unique dynamic weighting strategy that does not strictly adhere to a feature contribution hierarchy (as revealed by the results from dataset B). This behavior is likely attributable to its distinctive architecture, which is designed to be trained across datasets and infer from entire datasets rather than individual samples.

Unlike traditional models that follow a fit-then-predict approach, TabPFN incorporates in-context learning. This method integrates training and inference within a single forward pass, enhancing the algorithm's robustness in environments (14) with sparse data. The design of TabPFN thus offers a considerable advantage in medical and scientific domains, where data heterogeneity and shifts in distribution are common; it helps consistently achieve reliable results across varied sample sizes and feature dimensions.

Despite the advantages offered by TabPFN in terms of rapid inference and minimal hyperparameter tuning, there are some limitations. Its applicability is primarily restricted to small-scale tabular classification problems (14,26), and it lacks flexibility for regression or complex multi-class tasks (14,26,27). In addition, as a black-box model, TabPFN provides limited interpretability compared with traditional ML methods such as logistic regression or RFs (28); this potentially hinders its acceptance in clinical practice (29). Implementation may require adequate hardware resources and a moderate learning curve for integration into clinical settings (27,30). Additionally, TabPFN parameters are pre-trained and fixed; the model may therefore perform suboptimally when data significantly diverges from the distributions of synthetic datasets used during pre-training (14,26,27). Further studies are needed to assess its robustness and generalizability in high-dimensional, sparse, or noisy real-world datasets, and explore potential model enhancements to improve flexibility, interpretability, and scalability (27-29).

However, our study has certain limitations. The small size of the datasets, which focused on rare kidney conditions, precluded comprehensive evaluation of the TabPFN algorithm. Future studies need to include more prevalent cases (such as renal clear cell carcinoma and other benign renal tumors) and incorporate larger sample sizes for comprehensive assessment. Additionally, although TabPFN was compared to other major ML algorithms, constraints related to the length of the study prevented a thorough evaluation of all algorithms. Nevertheless, the results consistently indicated that TabPFN offered stability and robustness. Another limitation was the restricted scope of our research group, which prevented the inclusion of lesions from other organs. Nonetheless, we utilized data from previous studies to evaluate the TabPFN algorithm, aiming to provide more definitive results.

Conclusions

The TabPFN algorithm demonstrated outstanding and stable diagnostic performance in identifying benign and malignant kidney tumors. It can be particularly effective in scenarios involving small sample sizes and extensive radiomic features. TabPFN may, therefore, be poised to meet the demands of real-world clinical research, marking it as a potentially invaluable tool in the ML arsenal for medical applications.

Acknowledgments

The authors would like to thank Medjaden Inc. for the scientific editing of this manuscript.

Footnote

Reporting Checklist: The authors have completed the CLEAR reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1132/rc

Data Sharing Statement: Available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1132/dss

Funding: This research was funded by Shenzhen Science and Technology Program (No. JCYJ20240813114631042).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-2025-1132/coif). Z.S. is an employee of Philips Healthcare. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are

appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments. This study was approved by the Medical Ethics Committee of the Guangdong Provincial Hospital of Chinese Medicine (approval Nos. ZE2023-090-01 for dataset A and ZE2024-294-01 for dataset B). The Ethics Committee of the Provincial Hospital of Chinese Medicine waived the requirement for written informed consent because the study involved a retrospective analysis of de-identified imaging and clinical data, posed no more than minimal risk to participants, and it was impracticable to obtain consent from all individuals due to the time span of data collection. All datasets were anonymised before analysis and no additional procedures were performed. All participating institutions were also informed and agreed the study.

References

1. Noda R, Ichikawa D, Shibagaki Y. Machine learning-based diagnostic prediction of minimal change disease: model development study. Sci Rep 2024;14:23460.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563-77.
3. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform 2023;24:bbad002.
4. Lee KS, Park H. Machine learning on thyroid disease: a review. Front Biosci (Landmark Ed) 2022;27:101.
5. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep 2022;24:523-33.
6. Malashin I, Tynchenko V, Gantimurov A, Nelyub V, Borodulin A. Support Vector Machines in Polymer Science: A Review. Polymers (Basel) 2025;17:491.
7. Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, Schirrmeister RT, Hutter F. Accurate predictions on small data with a tabular foundation model. Nature 2025;637:319-26.
8. Araújo ALD, Moraes MC, Pérez-de-Oliveira ME, Silva VMD, Saldivia-Siracusa C, Pedroso CM, Lopes MA, Vargas PA, Kochanny S, Pearson A, Khurram SA, Kowalski LP, Migliorati CA, Santos-Silva AR. Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis. Oral Oncol 2023;140:106386.
9. Huang L, Ye Y, Chen J, Feng W, Peng S, Du X, Li X, Song Z, Liu T. Cystic renal mass screening: machine-learning-based radiomics on unenhanced computed tomography. Diagn Interv Radiol 2024;30:236-47.
10. Ye Y, Weng B, Guo Y, Huang L, Xie S, Zhong G, Feng W, Lin W, Song Z, Wang H, Liu T. Intratumoral and peritumoral radiomics using multi-phase contrast-enhanced CT for diagnosis of renal oncocytoma and chromophobe renal cell carcinoma: a multicenter retrospective study. Front Oncol 2025;15:1501084.
11. Huang L, Feng W, Lin W, Chen J, Peng S, Du X, Li X, Liu T, Ye Y. Enhanced and unenhanced: Radiomics models for discriminating between benign and malignant cystic renal masses on CT images: A multi-center study. PLoS One 2023;18:e0292110.
12. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30:1323-41.
13. Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. SIAM Review 2018;60:223-311.
14. Hollmann N, Müller S, Eggensperger K, Hutter F. TabPFN: a transformer that solves small tabular classification problems in a second. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.
15. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565-7.
16. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med 2016;4:218.
17. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris, France, August 22-27, 2010 Keynote, Invited and Contributed Papers. Heidelberg: Physica-Verlag HD; 2010:177-86.
18. Goodfellow I, Bengio Y, Courville A. Deep learning.

Cambridge: MIT Press; 2016.
19. Liu M, Chen L, Du X, Jin L, Shang M. Activated Gradients for Deep Neural Networks. IEEE Trans Neural Netw Learn Syst 2023;34:2156-68.
20. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36:1-13.
21. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-82.
22. Ziegel ER. The elements of statistical learning. Technometrics 2003;45:267-8.
23. Breiman L. Random forests. Machine Learning 2001;45:5-32.
24. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-94.
25. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017.
26. Müller S, Hollmann N, Arango SP, Grabocka J, Hutter F. Transformers can do Bayesian inference. In: Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022). 2022.
27. Borisov V, Leemann T, Sebler K, Haug J, Pawelczyk M, Kasneci G. Deep Neural Networks and Tabular Data: A Survey. IEEE Trans Neural Netw Learn Syst 2024;35:7499-519.
28. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell 2019;1:206-15.
29. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, Mooney C. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences 2021;11:5088.
30. Raschka S, Patterson J, Nolet C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information 2020;11:193.

Cite this article as: Liu T, Wang H, Guo Y, Ye Y, Weng B, Li X, Chen J, Xie S, Zhong G, Song Z, Huang L. Tabular prior-data fitted network in real-world CT radiomics: benign *vs.* malignant renal tumor classification. Quant Imaging Med Surg 2025;15(11):10847-10861. doi: 10.21037/qims-2025-1132

Supplementary

Table S1 Clinical and pathological characteristics of the patients in dataset A

Characteristics	Training set (n=162)			Validation set (n=45)		
	Benign (n=77)	Malignant (n=85)	P value	Benign (n=15)	Malignant (n=30)	P value
Age (years)	57.4±9.8	60.3±12.6	0.746	58.2±10.6	61.6±14.0	0.633
Gender			0.281			0.831
Male	41 (19.8)	37 (17.87)		9 (4.35)	17 (8.21)	
Female	36 (12.56)	48 (23.18)		6 (2.90)	13 (6.28)	
Location			0.311			0.831
Right kidney	41 (19.80)	53 (25.60)		6 (2.90)	13 (6.28)	
Left kidney	36 (17.39)	32 (15.46)		9 (4.35)	17 (8.21)	
Histological subtype			<0.0001			<0.0001
Simple kidney cyst	77 (37.19)	0 (0.00)		14 (6.76)	0 (0.00)	
Clear cell carcinoma	0 (0.00)	84 (40.5)		0 (0.00)	30 (14.50)	
Other	0 (0.00)	1 (0.4)		1 (0.48)	0 (0.00)	
Bosniak category			<0.0001			<0.0001
IIF	63 (30.43)	16 (7.73)		12 (5.80)	3 (1.45)	
III	14 (6.76)	21 (10.14)		3 (1.45)	10 (4.83)	
IV	0 (0.00)	48 (23.18)		0 (0.00)	17 (8.21)	

Data are expressed as the mean ± SD or frequency (constituent ratio). SD, standard deviation.

Table S2 Clinical and pathological characteristics of the patients in dataset B

Characteristics	Training set (n=65)			Validation set (n=27)		
	CRCC (n=36)	RO (n=29)	P value	CRCC (n=15)	RO (n=12)	P value
Age (years)	56.4±15.7	51.4±13.63	0.10	51.4±13.63	54.5±8.96	0.69
Gender			0.39			0.93
Male	16 (17.58)	11 (12.09)		6 (6.60)	5 (5.50)	
Female	20 (21.98)	18 (19.78)		9 (9.89)	7 (7.69)	
Location			0.62			0.42
Right kidney	17 (18.68)	16 (17.58)		8 (8.79)	5 (5.50)	
Left kidney	19 (20.88)	13 (14.29)		7 (7.69)	7 (7.69)	

Data are expressed as mean ± SD, median (interquartile range), or frequency (constituent ratio). CRCC, chromophobe renal cell carcinoma; RO, renal oncocytoma; SD, standard deviation.

https://dx.doi.org/10.21037/qims-2025-1132

Table S3 Radiomic features used for analysis of datasets A and B

Dataset A radiomic features
Unenhanced of dataset A (4 features)
Original_glcm_MaximumProbability_Unenhanced
Wavelet-LHH_firstorder_Median_Unenhanced
Wavelet-LLL_firstorder_90Percentile_Unenhanced
Wavelet-LLL_firstorder_Median_Unenhanced
CMP of dataset A (2 features)
Wavelet-LLL_firstorder_Median
Wavelet-LLL_firstorder_Uniformity
NP of dataset A (5 features)
Original_firstorder_Median
Wavelet-LLL_gldm_DependenceEntropy
Original_firstorder_Mean
Original_glcm_MaximumProbability
Original_gldm_DependenceEntropy
Dataset B radiomic features
CMP of dataset B (9 features)
Log-sigma-1-mm-3D_firstorder_Skewness_CMP
Log-sigma-1-mm-3D_glcm_Imc2_CMP
Log-sigma-1-mm-3D_glrlm_RunLengthNonUniformityNormalized_CMP
Log-sigma-2-mm-3D_glcm_ClusterShade_CMP
Log-sigma-3-mm-3D_glszm_SmallAreaEmphasis_CMP
Log-sigma-3-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis_CMP
Wavelet-HLL_gldm_DependenceEntropy_CMP
Wavelet-HLL_gldm_LargeDependenceEmphasis_CMP
Wavelet-LLL_firstorder_RobustMeanAbsoluteDeviation_CMP
NP of dataset B (3 features)
Log-sigma-3-mm-3D_glszm_SmallAreaEmphasis_NP
Wavelet-HLL_gldm_DependenceVariance_NP
Wavelet-HLH_glszm_SmallAreaHighGrayLevelEmphasis_NP
Fusion of dataset B (12 features)
12 rad features from the CMP of dataset B (9 features) and NP of dataset B (3 features)

CMP, corticomedullary phase; NP, nephrographic phase.

https://dx.doi.org/10.21037/qims-2025-1132

Table S4 P values and z scores from the DeLong test for pair-wise comparisons of different ML models in the validation set of dataset A

Unenhanced				CMP				NP			
Algorithm A	Algorithm B	z score	P value	Algorithm A	Algorithm B	z score	P value	Algorithm A	Algorithm B	z score	P value
SVM	SGD	1.31	0.19	SVM	SGD	−1.02	0.31	SVM	SGD	3.09*	<0.001*
SVM	KNN	0.71	0.48	SVM	KNN	1.42	0.16	SVM	KNN	0.75	0.45
SVM	RF	−0.27	0.78	SVM	RF	−0.07	0.94	SVM	RF	−0.22	0.83
SVM	XGBoost	1.28	0.20	SVM	XGBoost	0.82	0.41	SVM	XGBoost	1.68	0.09
SVM	LightGBM	0.86	0.39	SVM	LightGBM	0.49	0.62	SVM	LightGBM	−0.13	0.89
SVM	TabPFN	0.33	0.74	SVM	TabPFN	0.84	0.40	SVM	TabPFN	−0.70	0.48
SGD	KNN	−0.85	0.40	SGD	KNN	2.26*	0.02*	SGD	KNN	−2.90*	<0.001*
SGD	RF	−1.82	0.07	SGD	RF	1.69	0.09	SGD	RF	−3.25*	<0.001*
SGD	XGBoost	−0.31	0.76	SGD	XGBoost	2.00	0.05	SGD	XGBoost	−2.26*	0.02*
SGD	LightGBM	−0.89	0.37	SGD	LightGBM	2.13*	0.03*	SGD	LightGBM	−3.00*	<0.001*
SGD	TabPFN	−1.02	0.31	SGD	TabPFN	1.37	0.17	SGD	TabPFN	−3.35*	<0.001*
KNN	RF	−1.11	0.27	KNN	RF	−1.82	0.07	KNN	RF	−0.97	0.33
KNN	XGBoost	0.76	0.44	KNN	XGBoost	−1.43	0.15	KNN	XGBoost	1.46	0.14
KNN	LightGBM	0.25	0.80	KNN	LightGBM	−1.73	0.08	KNN	LightGBM	−0.72	0.47
KNN	TabPFN	−0.03	0.98	KNN	TabPFN	−1.04	0.30	KNN	TabPFN	−1.17	0.24
RF	XGBoost	1.74	0.08	RF	XGBoost	1.27	0.20	RF	XGBoost	1.85	0.06
RF	LightGBM	1.28	0.20	RF	LightGBM	1.25	0.21	RF	LightGBM	0.15	0.88
RF	TabPFN	0.59	0.56	RF	TabPFN	0.78	0.44	RF	TabPFN	−0.45	0.65
XGBoost	LightGBM	−0.71	0.47	XGBoost	LightGBM	−1.28	0.20	XGBoost	LightGBM	−1.55	0.12
XGBoost	TabPFN	−0.66	0.51	XGBoost	TabPFN	−0.21	0.83	XGBoost	TabPFN	−1.94	0.05
LightGBM	TabPFN	−0.25	0.80	LightGBM	TabPFN	0.21	0.83	LightGBM	TabPFN	−0.65	0.51

*, P<0.05. z score presented as positive when algorithm A was better than algorithm B, and z score presented as negative when algorithm A was worse than algorithm B. CMP, corticomedullary phase; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; NP, nephrographic phase; RF, random forest; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

https://dx.doi.org/10.21037/qims-2025-1132

Table S5 P values and z scores from the DeLong test for pair-wise comparisons of different ML models in the validation set of dataset B

CMP				NP				Fusion			
Algorithm A	Algorithm B	z score	P value	Algorithm A	Algorithm B	z score	P value	Algorithm A	Algorithm B	z score	P value
SVM	SGD	0.79	0.43	SVM	SGD	0.17	0.86	SVM	SGD	1.65	0.10
SVM	KNN	−0.39	0.70	SVM	KNN	−0.12	0.90	SVM	KNN	−0.10	0.92
SVM	RF	−1.25	0.21	SVM	RF	1.16	0.24	SVM	RF	−1.03	0.30
SVM	XGBoost	−0.49	0.63	SVM	XGBoost	1.83	0.07	SVM	XGBoost	−0.44	0.66
SVM	LightGBM	−0.50	0.62	SVM	LightGBM	1.60	0.11	SVM	LightGBM	0.14	0.89
SVM	TabPFN	−1.29	0.20	SVM	TabPFN	0.92	0.36	SVM	TabPFN	−0.98	0.33
SGD	KNN	−1.26	0.21	SGD	KNN	−0.30	0.76	SGD	KNN	−1.89	0.06
SGD	RF	−3.38*	<0.001*	SGD	RF	1.18	0.24	SGD	RF	−3.73*	<0.001*
SGD	XGBoost	−1.82	0.07	SGD	XGBoost	1.78	0.07	SGD	XGBoost	−2.50*	0.01*
SGD	LightGBM	−1.57	0.12	SGD	LightGBM	1.50	0.13	SGD	LightGBM	−1.61	0.11
SGD	TabPFN	−3.49*	<0.001*	SGD	TabPFN	1.05	0.29	SGD	TabPFN	−3.06*	<0.001*
KNN	RF	−1.06	0.29	KNN	RF	1.18	0.24	KNN	RF	−1.01	0.31
KNN	XGBoost	−0.33	0.74	KNN	XGBoost	2.67*	0.01*	KNN	XGBoost	−0.37	0.71
KNN	LightGBM	−0.26	0.79	KNN	LightGBM	2.33*	0.02*	KNN	LightGBM	0.22	0.83
KNN	TabPFN	−1.12	0.26	KNN	TabPFN	1.03	0.30	KNN	TabPFN	−0.80	0.42
RF	XGBoost	1.42	0.15	RF	XGBoost	1.11	0.27	RF	XGBoost	1.20	0.23
RF	LightGBM	1.33	0.18	RF	LightGBM	1.16	0.24	RF	LightGBM	2.16*	0.03*
RF	TabPFN	−0.11	0.91	RF	TabPFN	−0.38	0.70	RF	TabPFN	0.52	0.60
XGBoost	LightGBM	0.11	0.91	XGBoost	LightGBM	0.38	0.70	XGBoost	LightGBM	1.12	0.26
XGBoost	TabPFN	−1.40	0.16	XGBoost	TabPFN	−1.17	0.24	XGBoost	TabPFN	−0.64	0.52
LightGBM	TabPFN	−1.46	0.14	LightGBM	TabPFN	−1.21	0.23	LightGBM	TabPFN	−1.54	0.12

*, P<0.05. z score presented as positive when algorithm A was better than algorithm B, and z score presented as negative when algorithm A was worse than algorithm B. CMP, corticomedullary phase; KNN, k-nearest neighbor; LightGBM, light gradient boosting machine; ML, machine learning; NP, nephrographic phase; RF, random forest; SGD, stochastic gradient descent; SVM, support vector machine; TabPFN, tabular prior-data fitted network; XGBoost, extreme gradient boosting.

https://dx.doi.org/10.21037/qims-2025-1132

Figure S1 Feature importance bar chart of ML classifiers in unenhanced, CMP, and NP groups of dataset A. CMP, corticomedullary phase; ML, machine learning; NP, nephrographic phase.

Figure S2 Feature importance bar chart of ML classifiers in unenhanced, CMP, and NP groups of dataset B. CMP, corticomedullary phase; ML, machine learning; NP, nephrographic phase.

https://dx.doi.org/10.21037/qims-2025-1132