Supplementary table S1 Patient and tumor characteristics in training, validation and test sets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Training set (n = 816) | Validation set(n = 234) | Test set(n = 112) | P values |
| Age median, range (years)GenderMale  FemaleTumor size, median, range (cm)Laterality Left RightLocation Upper Interpole LowerApproach for diagnosis  Surgery excision Biopsy ImagingTumor subtype | 56.0 (5‒88)404 (49.5%)412 (50.5%)2.7 (0.2‒18.1)407 (49.9%)409 (50.1%)267 (32.7%)328 (40.2%)221 (27.1%)489 (59.9%)12 (1.5%)315 (38.6%) | 59.5 (7‒92)120 (51.3%)114 (48.7%)2.6 (0.6‒14.6)112 (47.9%)122 (52.1%)75 (32.1%)94 (40.2%)65 (27.8%)199 (95.0%)5 (2.2%)30 (12.8%) | 63.0(17‒90)68 (60.7%)44 (39.3%)3.1 (0.6‒18.7)52 (45.5%)64 (54.5%)41 (36.6%)43 (38.4%)28 (25.0%)110 (98.2%)2 (1.8%)0 (0%) | < 0.001\*0.0840.1050.6350.935< 0.001\*< 0.001\* |
|  Clear cell RCC Papillary RCC Chromophobe RCCClear cell papillary RCCMultilocular cystic RCCUnclassified RCCOncocytomaAngiomyolipomaMixed epithelial and stromal tumorMetanephric adenomaRenal adenomaFuhrman/ ISUP gradeGrade 1Grade 2Grade 3Grade 4UnavailableInstitution HUP SXY PHH MAY TCIA | 292 (35.8%)113 (13.8%)24 (2.9%)18 (2.2%)7 (0.9%)6 (0.7%) 37 (4.5%)318 (39.0%)0 (0%)0 (0%)1 (0.1%)49 (10.1%)212 (43.9%)109 (22.6%)26 (5.4%)87 (18.0%)673 (82.4%)19 (2.3%)23 (2.8%)62 (7.6%)39 (4.8%) | 86 (36.8%)31 (13.2%)6 (2.6%)7 (3.0%)0 (0%)2 (0.9%)21 (9.0%)74 (31.6%)2 (0.9%)3 (1.3%)2 (0.9%)12 (7.0%)65 (38.0%)35 (20.5%)7 (4.1%)52 (30.4%)174 (74.4%)3 (1.3%)18 (6.8%)39 (14.1%)8 (3.4%) | 47 (42.0%)8 (7.1%)2 (1.8%)5 (4.5%)0 (0%)1 (0.9%)34 (30.4%)14 (12.5%)1 (0.9%)0 (0%)0 (0%)7 (6.8%)30 (29.1%)16 (15.5%)2 (1.9%)48 (46.6%)66 (58.9%)3 (2.7%)11 (9.8%)23 (20.5%)9 (8.0%) | < 0.001\*< 0.001\* |

RCC, renal cell carcinoma; ISUP, International Society of Urological Pathology;

\* Statistically significant

Supplementary table S2. Ensemble cross-validation test set performance

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fold | F1 Score | ROC AUC | PR AUC | Acc (95% CI) | TPR (95% CI) | TNR (95% CI) | PPV | NPV | FDR |
| 1 | 0.76  | 0.59  | 0.85  | 0.65 (0.56-0.72) | 0.68 (0.59-0.76) | 0.48 (0.30-0.67) | 0.86  | 0.25  | 0.14  |
| 2 | 0.79  | 0.67  | 0.83  | 0.70 (0.62-0.76) | 0.75 (0.67-0.82) | 0.54 (0.39-0.68) | 0.82  | 0.44  | 0.18  |
| 3 | 0.75  | 0.60  | 0.86  | 0.64 (0.56-0.71) | 0.68 (0.59-0.76) | 0.46 (0.30-0.64) | 0.84  | 0.27  | 0.16  |
| 4 | 0.72  | 0.56  | 0.83  | 0.60 (0.52-0.68) | 0.64 (0.55-0.72) | 0.44 (0.28-0.63) | 0.83  | 0.23  | 0.17  |
| 5 | 0.75  | 0.61  | 0.84  | 0.64 (0.56-0.71) | 0.70 (0.61-0.77) | 0.44 (0.29-0.61) | 0.81  | 0.30  | 0.19  |
| Average | 0.75  | 0.60  | 0.84  | 0.64 (0.56-0.72) | 0.69 (0.60-0.77) | 0.47 (0.31-0.65) | 0.83  | 0.30  | 0.17 |

ROC AUC, Area under Receiver Operating Characteristic curve; PR AUC, Area under Precision-Recall curve; TPR, True positive rate or sensitivity; TNR True negative rate or specificity; Acc, Accuracy; PPV, Positive predictive value; NPV, Negative predictive value; FDR, False discovery rate; N/A, Not applicable; 95% CI, 95% confidence interval.

Supplementary table S3. Performance of T1C, T2, clinical features and ensemble models in independent test set

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Modality | F1 Score | ROC AUC | PR AUC | Acc (95% CI) | TPR (95% CI) | TNR (95% CI) | PPV | NPV | FDR |
| Clinical | 0.72 | 0.44 | 0.52 | 0.57 (0.48-0.66) | 0.97 (0.89-1.00) | 0.08 (0.02-0.18) | 0.57 | 0.67 | 0.43 |
| T1C | 0.61 | 0.53 | 0.57 | 0.55 (0.46-0.63) | 0.65 (0.53-0.76) | 0.42 (0.29-0.55) | 0.58 | 0.49 | 0.42 |
| T2 | 0.74 | 0.61 | 0.63 | 0.67 (0.58-0.75) | 0.85 (0.74-0.92) | 0.45 (0.33-0.59) | 0.66 | 0.71 | 0.34 |
| Ensemble | 0.75 | 0.60 | 0.66 | 0.68 (0.59-0.76) | 0.86 (0.76-0.93) | 0.45 (0.33-0.59) | 0.66 | 0.73 | 0.34 |

ROC AUC, Area under Receiver Operating Characteristic curve; PR AUC, Area under Precision-Recall curve; TPR, True positive rate or sensitivity; TNR True negative rate or specificity; Acc, Accuracy; PPV, Positive predictive value; NPV, Negative predictive value; FDR, False discovery rate; N/A, Not applicable; 95% CI, 95% confidence interval.

Supplementary table S4. Median, mean, standard deviation, and relative standard

deviation percent of each classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Median AUC | Mean AUC | St. Deviation | RSD% |
| GLM | 0.54 | 0.54 | 0.04 | 1.92 |
| NNet | 0.53 | 0.53 | 0.03 | 4.99 |
| KNN | 0.54 | 0.54 | 0.02 | 2.43 |
| BST | 0.53 | 0.53 | 0.02 | 1.96 |
| BAG | 0.54 | 0.54 | 0.02 | 2.41 |
| RF | 0.53 | 0.53 | 0.03 | 2.21 |
| LDA | 0.54 | 0.55 | 0.03 | 2.09 |
| BY | 0.51 | 0.52 | 0.02 | 3.96 |
| DT | 0.52 | 0.52 | 0.02 | 2.90 |
| SVM | 0.50 | 0.50 | 0.01 | 0.207 |

GLM, Generalized Linear Models; NNet, Neural Network; KNN, K-Nearest-Neighbor; BST, Binary Search Tree; BAG, Bagging; RF, Random Forest; LDA, Linear Discrimant Analysis; BY, Bayes; DT, Decision Tree; SVM, Support Vector Machine; AUC, Area under curve; RSD; Relative standard deviation.

Supplementary table S5. The test performances of the best hand-optimized machine learning classifier and the TPOT pipeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | AUC | Accuracy | Sensitivity | Specificity |
| LDA | 0.61 | 0.62 | 0.76 | 0.46 |
| TPOT | 0.58 | 0.61 | 0.80 | 0.37 |

LDA, Linear Discrimant Analysis; Tree-Based Pipeline Optimization Tool

Figure legends

Supplementary Figure S1. Flow diagram of patient inclusion and exclusion.

Supplementary Figure S2. Heatmap of validation performances (in AUC) of hand-optimized machine learning classifiers.

Supplementary Figure S3. Confusion matrices for all models across training (A), validation (B), and test cohorts (C).

Supplementary Figure S4. Reliability curve of all models across training (A), validation (B), and test cohorts (C).

Supplementary Figure S5. The pipeline exported by TPOT Auto-ML script.

Supplementary Figure S6. Reliability diagram of the ensemble model across training (A), validation (B), and test cohorts (C).