Supplementary information of https://doi.org/10.1631/jzus.B2500607

Materials and methods; Figs. S1–S3; Tables S1–S3

Materials and methods

Detection of PRKD3 expression levels

The enzyme-linked immunosorbent assay (ELISA) kit used was the "Research-Grade Human protein kinase D3(PRKD3) Protein Kit" (catalog number YJ955112) purchased from Shanghai Yuanju Biotechnology Center. The detection range was 1–80 μg/mL, and the limit of detection (LOD) was 1.0 μg/mL. The inter-plate coefficient of variation (CV) ranged from 3% to 10%, and the intra-plate CV ranged from 2.5% to 7.0%. The intra-assay precision CV was ≤4.5%, and the inter-assay precision CV was ≤6.0%. In the ELISA procedure, both the test samples and standards were assayed in duplicate wells. For each microtiter plate, six standard solutions with serial concentrations were added to generate the standard curve. After the assay, the mean absorbance (optical density (OD)) values of the duplicate wells for each standard concentration were calculated, and these values were subsequently used to fit and establish the standard curve. The assay was considered valid only when the correlation coefficient (R) of the linear regression between the standard concentrations and corresponding absorbance values was ≥ 0.9900. Meanwhile, the CV of absorbance values across all duplicate wells was required to be <10% to ensure assay stability and reproducibility. In addition, blank wells and negative control wells were included in each microtiter plate, and normal saline was added to the negative control wells. If a sample's detected value exceeded the kit's detection range, the sample was subjected to serial dilution and re-assayed. If samples could not be tested immediately after collection, they were stored at −20 °C or below to avoid repeated freeze–thaw cycles.

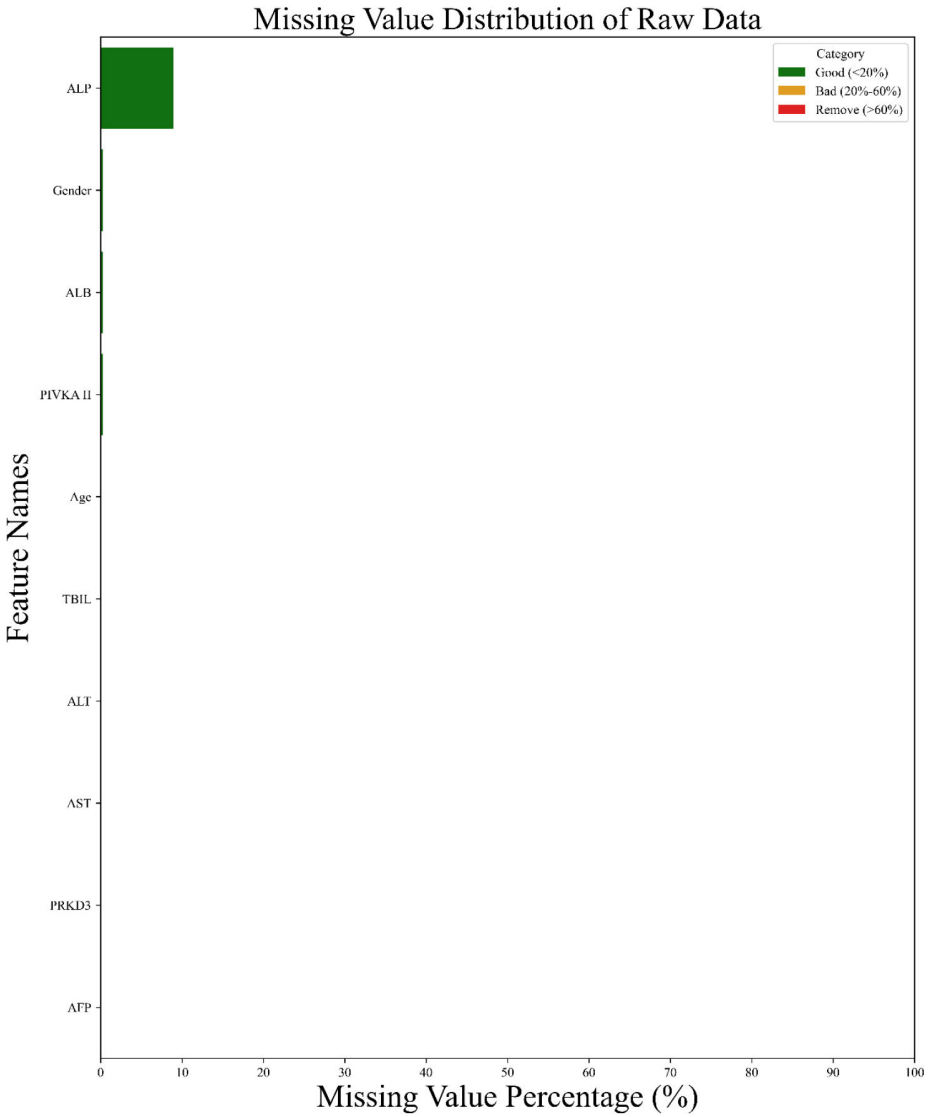


Fig. S1 Overview of missing values in the raw dataset prior to any processing.

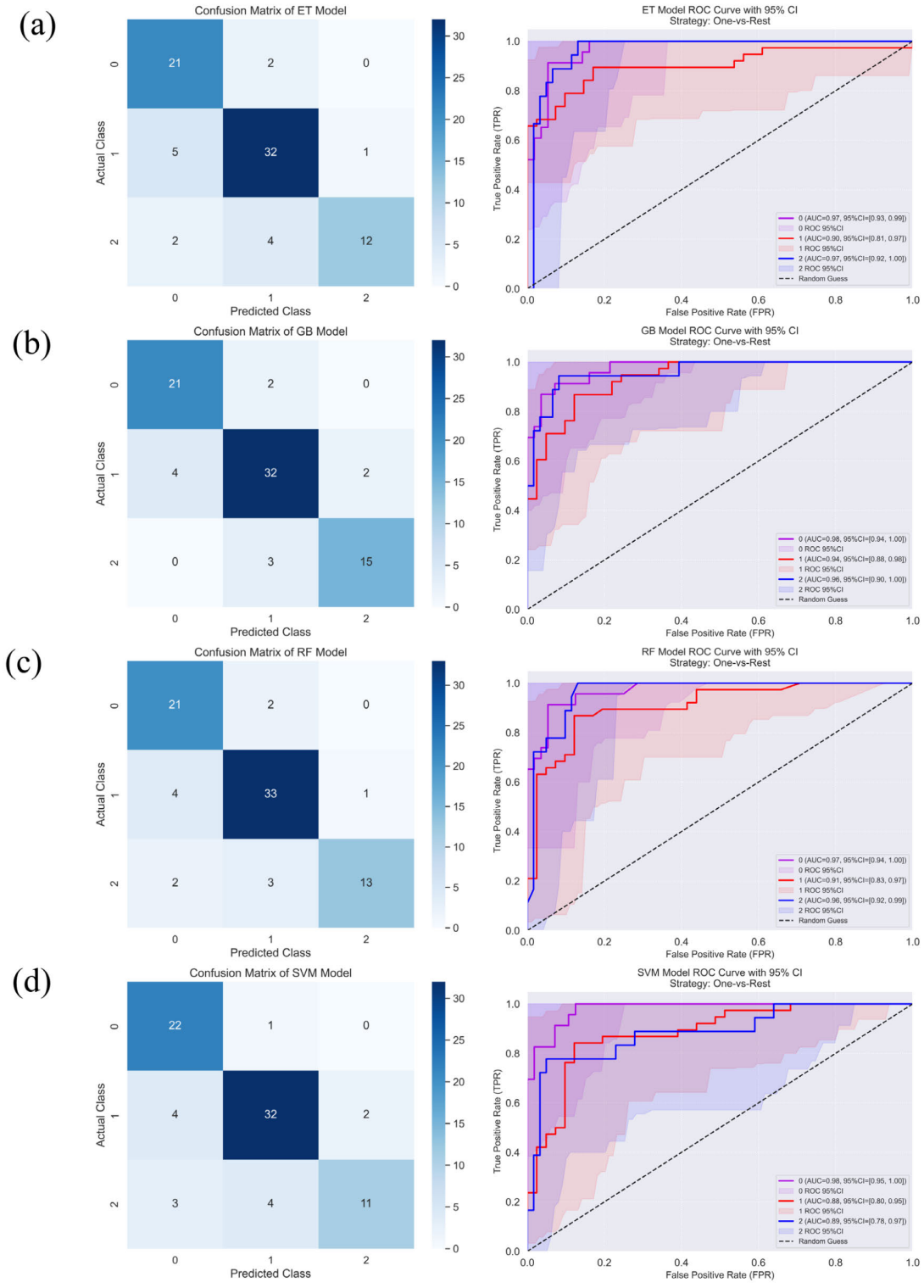


Fig. S2 Confusion matrix and ROC curve for each base model. (a) Confusion matrix and ROC curve with 95% CI of the extra trees (ET) model. (b) Confusion matrix and ROC curve with 95% CI of the gradient boosting (GB) model. (c) Confusion matrix and ROC curve with 95% CI of the random forest (RF) model. (d) Confusion matrix and ROC curve with 95% CI of the support vector machine (SVM) model.

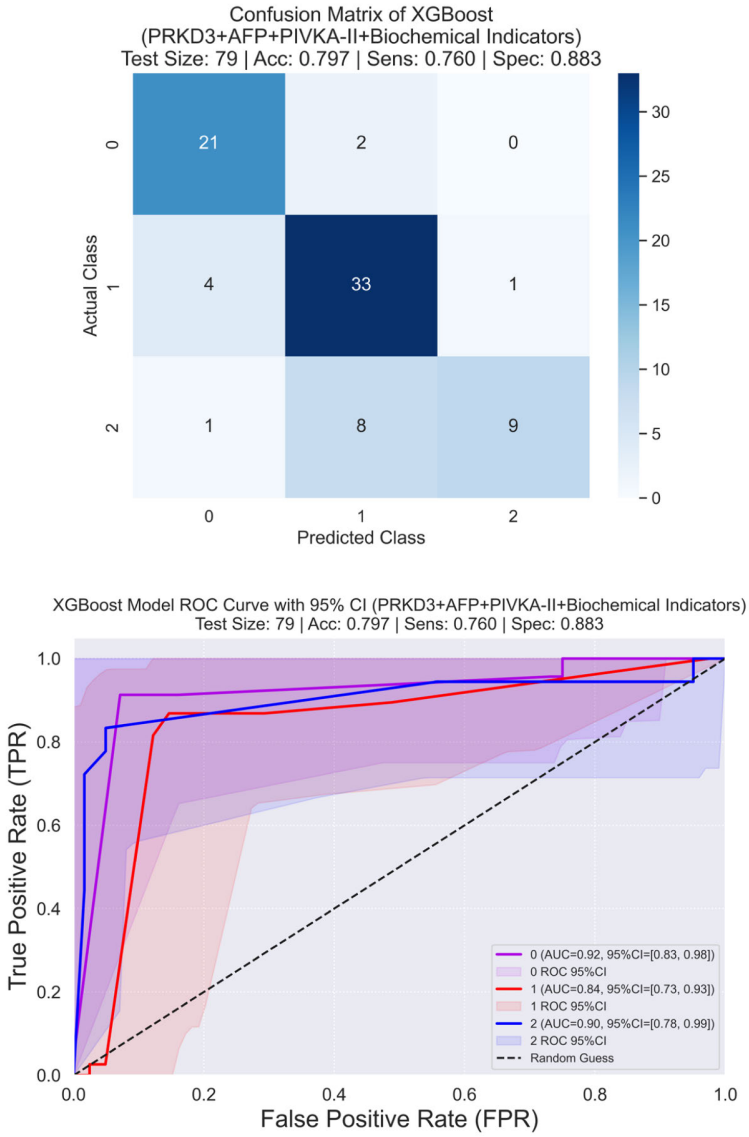


Fig. S3 Confusion matrix and ROC curve of the classifier on the original imbalanced test set (without SMOTE).

Table S1 Hyperparameter settings of key machine learning models

Model category	Model abbreviation	Core hyperparameter configuration	Random seed
Extra Trees Classifier	ET	*n*_estimators=200; max_depth=12; min_samples_leaf=1; min_samples_split=2; max_features=None; all other parameters follow library defaults	0
Gradient Boosting Classifier	GB	min_samples_leaf=4; min_samples_split=2; max_depth=3; learning_rate=0.1; *n*_estimators=160; subsample=1.0; all other parameters follow library defaults	0
Random Forest Classifier	RF	*n*_estimators=200; max_depth=None; max_features='log2'; min_samples_leaf=1; min_samples_split=2; criterion='gini'; all other parameters follow library defaults	0
Support Vector Machine	SVM	*C*=10; gamma='scale'; kernel='rbf'; all other parameters default; all other parameters follow library defaults	0
XGBoost Classifier	XGBoost	*n*_estimators=200; max_depth=3; learning_rate=0.2; colsample_bytree=1.0; reg_lambda=5; subsample=1.0; all others default; all other parameters follow library defaults	0

Software environment: Python 3.11.8, scikit-learn 1.2.2, XGBoost 2.0.3, imbalanced-learn 0.12.4, Pandas 2.1.4, NumPy 1.26.4, Matplotlib 3.8.0, Seaborn 0.12.2

Table S2 Comparison of test set performance between the SMOTE-balanced dataset and the original unbalanced dataset

Evaluation dataset type	ACC	SENS	SPEC	PREC	F1 score
SMOTE-balanced test set	0.861	0.863	0.925	0.862	0.861
Original unbalanced test set	0.797	0.760	0.883	0.809	0.788

SMOTE: synthetic minority oversampling technique; ACC: accuracy; SENS: sensitivity; SPEC: specificity; PREC: precision.

Table S3 Summary of 4-repeat random split validation results

Random seed	ACC	F1 score	Healthy AUC (95% CI)	Cirrhosis or decompensated cirrhosis AUC (95% CI)	HCC AUC (95% CI)
0	0.861	0.861	0.944 (0.884, 0.987)	0.864 (0.776, 0.941)	0.935 (0.865, 0.989)
80	0.835	0.828	0.921 (0.830, 0.992)	0.883 (0.811, 0.964)	0.894 (0.861, 0.995)
123	0.823	0.909	0.961 (0.906, 1.000)	0.941 (0.883, 0.986)	0.831 (0.676, 0.952)
456	0.899	0.897	0.985 (0.959, 1.000)	0.949 (0.885, 0.994)	0.863 (0.690, 0.988)
Mean±SD	0.855±0.034	0.874±0.037	0.953±0.027	0.909±0.042	0.881±0.044

ACC: accuracy; AUC: area under the curve; CI: confidence interval; HCC: hepatocellular carcinoma; SD: standard deviation.