Clinical and Experimental Medicine (2025) 25:275
https://doi.org/10.1007/s10238-025-01818-5

RESEARCH

Innovative machine learning approach for liver fibrosis and disease severity evaluation in MAFLD patients using MRI fat content analysis

Mengting Hou[1] · Yujie Zhu[1] · Huadi Zhou[1] · Siyi Zhou[1] · Jianjun Zhang[1] · Yue Zhang[1] · Xiao Liu[1]

Received: 7 May 2025 / Accepted: 13 July 2025

Abstract

This study employed machine learning models to quantitatively analyze liver fat content from MRI images for the evaluation of liver fibrosis and disease severity in patients with metabolic dysfunction-associated fatty liver disease (MAFLD). A total of 26 confirmed MAFLD cases, along with MRI image sequences obtained from public repositories, were included to perform a comprehensive assessment. Radiomics features—such as contrast, correlation, homogeneity, energy, and entropy—were extracted and used to construct a random forest classification model with optimized hyperparameters. The model achieved outstanding performance, with an accuracy of 96.8%, sensitivity of 95.7%, specificity of 97.8%, and an F1-score of 96.8%, demonstrating its strong capability in accurately evaluating the degree of liver fibrosis and overall disease severity in MAFLD patients. The integration of machine learning with MRI-based analysis offers a promising approach to enhancing clinical decision-making and guiding treatment strategies, underscoring the potential of advanced technologies to improve diagnostic precision and disease management in MAFLD.

Keywords Magnetic resonance imaging · Machine learning · Liver fibrosis · Metabolic dysfunction-associated fatty liver disease · Radiomics · Quantitative analysis

Introduction

Metabolic dysfunction-associated fatty liver disease (MAFLD) is one of the most common chronic liver diseases worldwide [1–3]. In recent years, the incidence of MAFLD has risen significantly due to the increasing prevalence of obesity and metabolic syndrome [4–6]. The pathological spectrum of MAFLD ranges from simple steatosis (NAFL) to non-alcoholic steatohepatitis (NASH), which can ultimately progress to liver fibrosis, cirrhosis, and even hepatocellular carcinoma (HCC) [7–9]. MAFLD not only impairs patients' quality of life, but also increases the risk of liver-related complications and all-cause mortality [10–12]. Liver biopsy is currently the gold standard for evaluating liver fibrosis. However, it is invasive and limited in routine clinical practice due to sampling variability, procedure-related risks, and low patient compliance [13–15]. Consequently, there is an urgent need for noninvasive, accurate, and efficient imaging methods to quantitatively assess liver fibrosis in MAFLD patients and improve clinical decision-making and disease management.

Magnetic resonance imaging (MRI), as a noninvasive imaging modality, has been widely utilized in recent years for the diagnosis and evaluation of liver diseases [16–18]. Multi-sequence MRI techniques—including T1, T2, in-phase, and out-of-phase imaging—can provide detailed information on liver tissue structure and fat content [19, 20]. However, conventional radiomics approaches often rely on manual feature extraction and basic statistical analyses, which are time-consuming, labor-intensive, and prone to observer bias, leading to limited reproducibility and reliability [21–23]. Furthermore, traditional methods struggle with the analysis of high-dimensional, complex imaging data, resulting in suboptimal predictive performance and generalizability [24]. To address these challenges, advanced machine learning (ML) techniques—particularly deep learning—offer promising solutions for automated image analysis and feature extraction [25].

✉ Yue Zhang
1264072839@qq.com

✉ Xiao Liu
cannelyes@163.com

[1] Department of Radiology, Zhejiang Hospital, Hangzhou, Zhejiang Province, China

Published online: 05 August 2025

ML approaches have demonstrated considerable success in various medical imaging tasks by learning discriminative features from large datasets to support disease prediction and classification [26–28]. Among them, the random forest (RF) algorithm is a widely used ensemble learning method known for its robustness in handling high-dimensional data and its effectiveness in feature selection [29–31]. By building multiple decision trees and aggregating their predictions, the RF model enhances classification stability and accuracy [32, 33]. In addition, RF models offer good interpretability, allowing identification of the most informative features contributing to classification outcomes [34, 35]. However, despite these advantages, research applying ML—particularly RF—to evaluate liver fibrosis in MAFLD remains limited and requires further validation [36, 37].

In this study, MRI data from MAFLD patients were collected from publicly available databases, including the Cancer Imaging Archive and Liver Imaging Database. The dataset included four imaging sequences: T1, T2, in-phase, and out-of-phase. Based on extracted radiomics features, we constructed training and validation datasets to build a classification model using the RF algorithm. To enhance performance, key hyperparameters—n_estimators and max_depth—were optimized through tenfold cross-validation. The model's classification results were then compared with pathological examination findings to evaluate diagnostic performance, identify limitations, and guide further model refinement.

The primary aim of this study is to quantitatively assess liver fat content using ML-based analysis of MRI images, in order to evaluate the degree of liver fibrosis and disease severity in MAFLD patients. Specifically, we sought to develop an efficient RF model capable of extracting informative features from multimodal MRI data for accurate disease classification. The results demonstrated that the proposed model achieved high accuracy, sensitivity, and specificity in assessing liver fibrosis, with performance closely aligned with that of pathological assessments. These findings suggest that ML-based MRI analysis may serve as a valuable noninvasive tool to support clinical diagnosis and therapeutic decision-making in MAFLD, potentially reducing the reliance on invasive procedures. Furthermore, this study provides preliminary evidence and technical support for future large-scale, multicenter research, contributing to the advancement of personalized and precise management of MAFLD.

Materials and methods

Collection of MRI imaging data

In this study, MRI data of patients diagnosed with MAFLD were retrieved from publicly accessible repositories, primarily the Cancer Imaging Archive (https://www.cancerimagingarchive.net/) and the Liver Imaging Database (https://liveratlas.org/). All MRI scans were acquired using 1.5 Tesla scanners, with an original matrix resolution of 256 × 256 pixels per image.

Each patient included in the dataset had a complete set of four MRI sequences: T1-weighted, T2-weighted, in-phase, and opposed-phase images. These sequences were chosen due to their common clinical usage in assessing hepatic steatosis, fibrosis, and tissue heterogeneity. Specifically, the in-phase and opposed-phase sequences are highly sensitive to hepatic fat content, the T2-weighted sequence reflects variations in tissue water content, and the T1-weighted sequence enhances tissue contrast. Together, these sequences provide complementary information, improving the robustness of radiomics-based feature extraction.

In some cases, corresponding histopathological reports obtained through percutaneous liver biopsy were available. Based on database annotations, the degree of liver fibrosis was categorized using a binary classification system: low vs. moderate-to-high fibrosis. This labeling was primarily derived from FibroScan liver stiffness measurements and subsequently mapped to the World Health Organization (WHO) grading criteria, where Mild fibrosis = F0–F1 and Moderate-to-severe fibrosis = F2–F4.

To support comprehensive analysis, additional clinical information was collected, including patient medical history, laboratory test results, and pathology reports. A rigorous data cleaning and annotation process was implemented to ensure dataset integrity and consistency. Samples were excluded if they exhibited any of the following: motion artifacts, excessive image noise or distortion, and missing essential clinical or imaging data. Liver region annotation in the MRI images was conducted by experienced radiologists, ensuring precise region-of-interest (ROI) delineation for subsequent feature extraction and model training (Fig. 1).

Note: The study downloaded imaging data of 26 different liver fibrosis patients from public datasets, including T1, T2, In_Phase, and Op_Phase four imaging states (as shown in the four sequences in the figure above). The data underwent preprocessing, feature extraction, feature selection, and model construction to obtain a complete fibrosis assessment of the patients, mainly classified as low fibrosis and high fibrosis.

Data preprocessing

During the preprocessing stage, all MRI images underwent a series of standardization procedures, including contrast enhancement and resizing to a uniform resolution of 256 × 256 pixels. To expand the dataset and improve model generalizability, data augmentation techniques were applied, including image rotation, horizontal and vertical flipping,

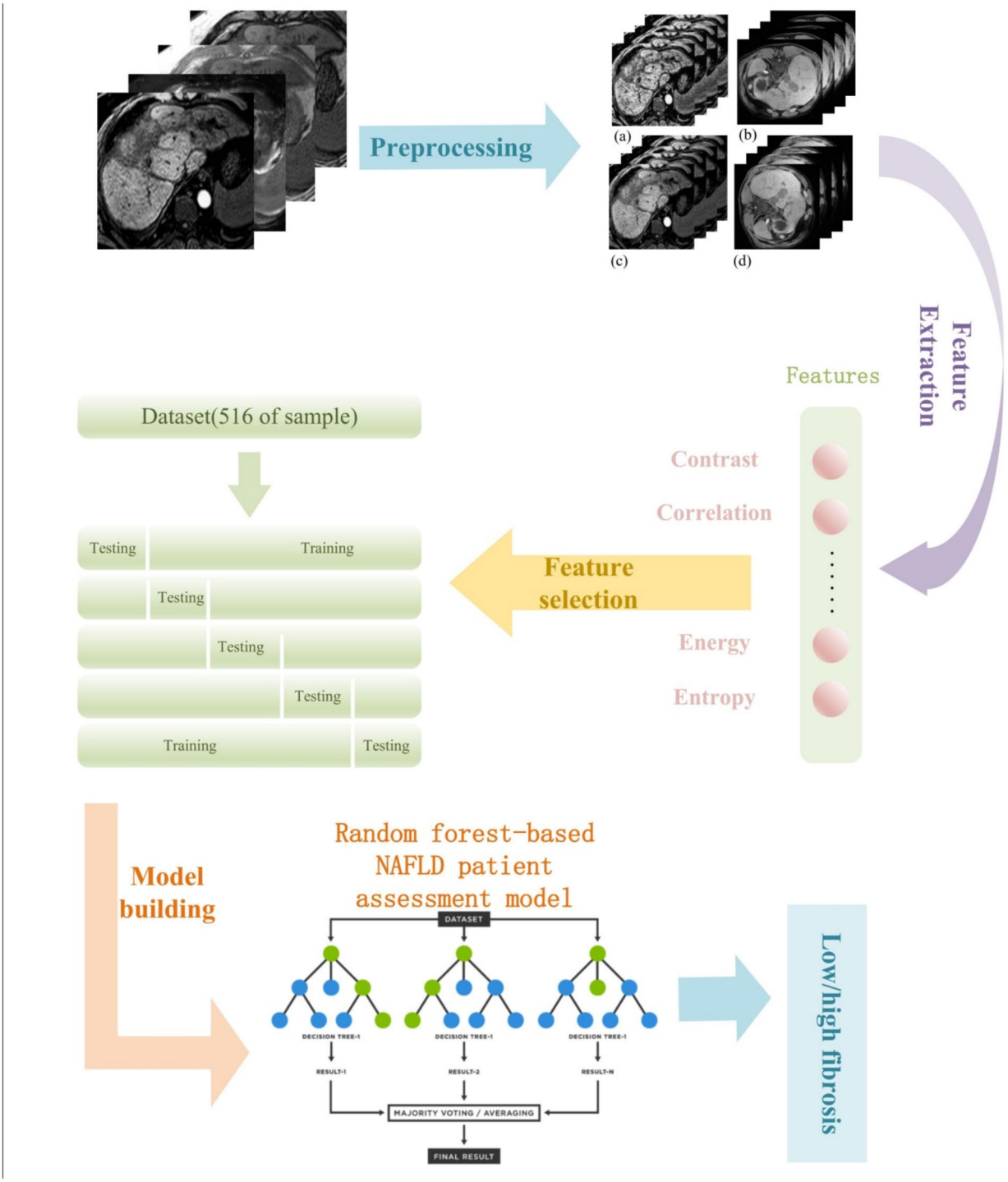


Fig. 1 Schematic diagram of the research process

and random cropping. Following preprocessing and augmentation, a total of 312 high-quality images were obtained for model development and evaluation (Fig. 2), effectively reducing the risk of overfitting.

For image denoising, a 2D Gaussian filter was applied using the cv2.GaussianBlur() function in Python, with a kernel size of (5, 5) and a standard deviation (σ) of 1.2. Histogram equalization was performed via cv2.equalizeHist() to improve contrast, gamma correction was applied with a gamma value of 1.5, and logarithmic transformation was implemented using a standard log mapping function to further enhance feature visibility. All images were

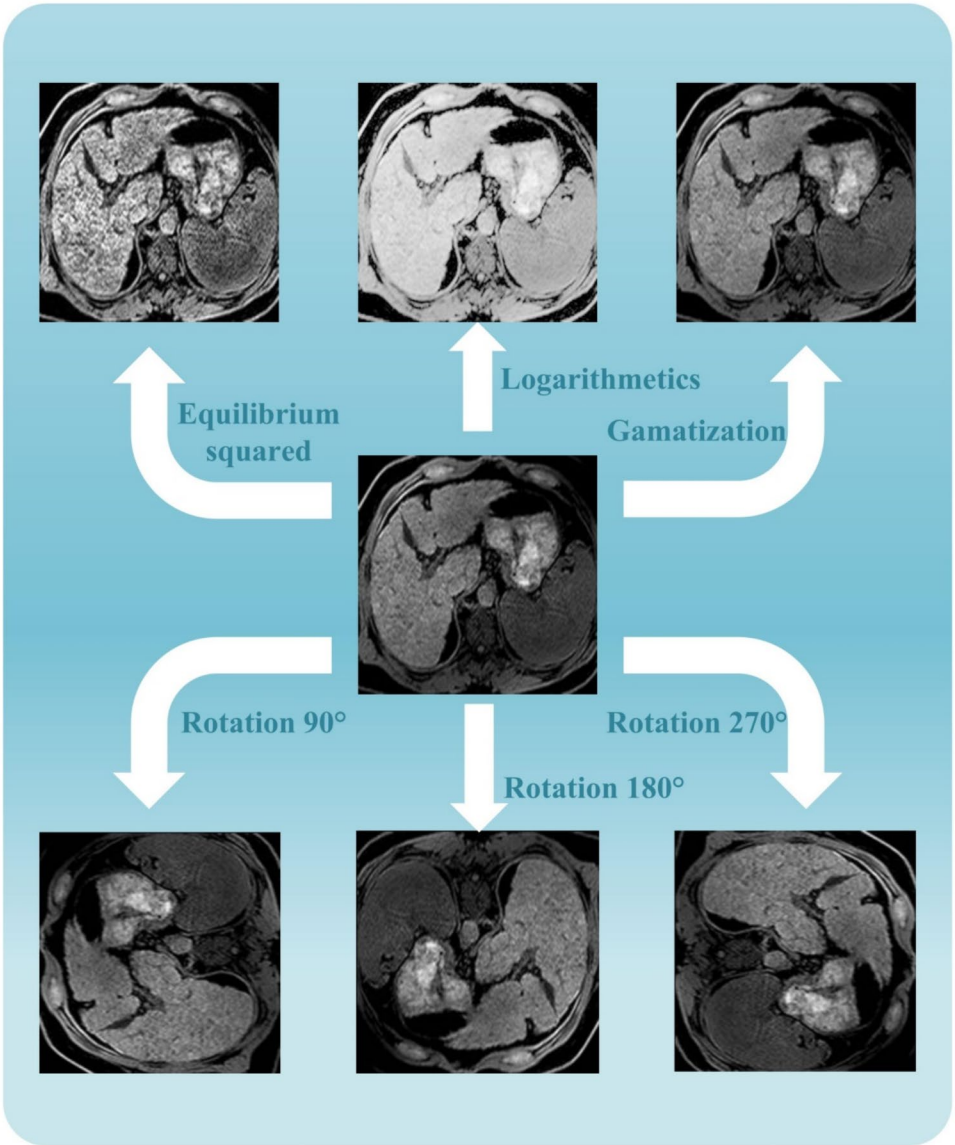


Fig. 2 Schematic diagram of data augmentation techniques for preprocessing MRI imaging data

subsequently normalized to a 0–1 scale and converted to grayscale with 255 Gy levels, thereby facilitating robust and consistent feature extraction. Collectively, these preprocessing steps—combined with data augmentation—substantially improved dataset diversity and enhanced the robustness of model training.

Feature extraction and selection

Feature extraction and selection were critical steps in constructing an effective ML model. Texture analysis was performed using gray-level co-occurrence matrices (GLCM) to extract five key radiomic features associated with hepatic fat distribution and fibrosis: contrast, correlation, homogeneity, energy, and entropy. These features are known to capture subtle structural heterogeneities in liver tissue that correlate with fat infiltration and fibrotic changes.

Given the potentially high dimensionality of the extracted features, feature selection was employed to reduce model complexity and enhance performance. The recursive feature elimination (RFE) method was used to iteratively train the model while removing the least informative features in each iteration. This process was repeated until the most representative features remained. As a result, the initial 20 extracted features were reduced to 10 selected features deemed most relevant for classification tasks.

Additionally, feature importance was assessed using the RF model, which calculates importance scores based on the mean decrease in Gini impurity (information gain) across all decision trees in the ensemble. These scores provided valuable insights into which features contributed most to

the model's classification performance and also improved interpretability.

Model selection and fine-tuning in ML

Following feature selection, the RF algorithm was selected as the primary ML classifier for model development (Fig. 3). RF enhances classification and regression accuracy by constructing an ensemble of decision trees and aggregating their predictive outcomes. Its notable advantages include the ability to model complex, nonlinear relationships in high-dimensional data, robustness against overfitting, efficient handling of missing values and imbalanced datasets, and relatively low sensitivity to hyperparameter tuning. Furthermore, the RF algorithm offers feature importance scoring, which enhances model interpretability and aids in understanding the decision-making process. After model selection, hyperparameter optimization was performed to improve predictive performance and generalizability. We employed grid search in conjunction with k-fold cross-validation to systematically explore hyperparameter combinations and evaluate model performance. Grid search exhaustively tests predefined parameter ranges, while cross-validation divides the dataset into training and validation subsets in iterative cycles to ensure robustness. The two most influential hyperparameters—n_estimators (number of trees in the forest) and max_depth (maximum depth of each tree)—were fine-tuned and optimized to n_estimators = 43 and max_depth = 6, which yielded the best model performance in our experiments.

While additional hyperparameters such as min_samples_split, min_samples_leaf, and max_features were also explored through preliminary tuning, their contributions to model performance were marginal. To maintain model simplicity and interpretability, we retained their default values and focused optimization efforts on the two most impactful parameters.

In parallel, we conducted preliminary comparisons with alternative classification models, including logistic regression (LR) and convolutional neural networks (CNNs). LR showed inferior performance in this small-sample, high-dimensional feature context, achieving an area under the curve (AUC) of approximately 0.85. CNNs, although

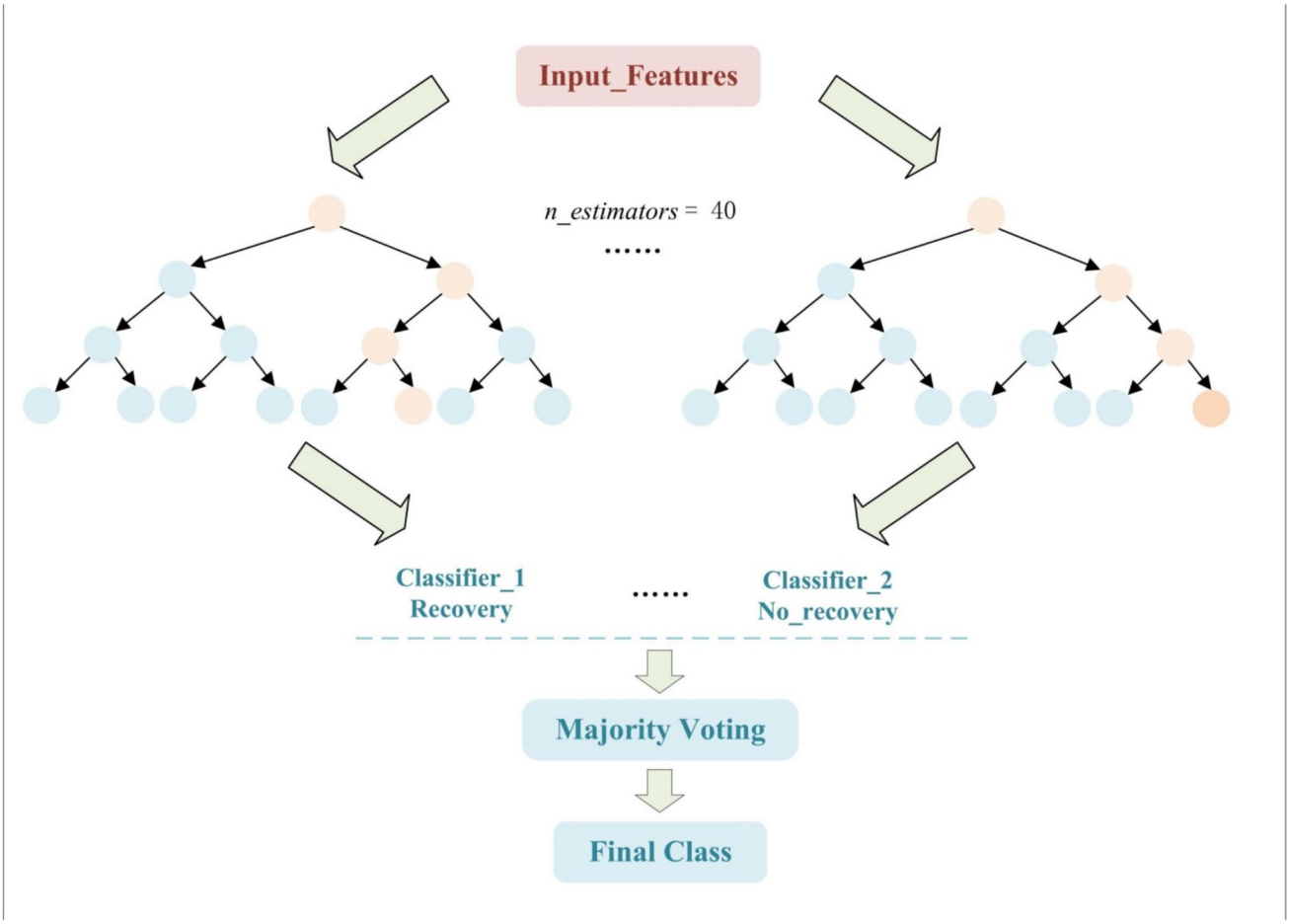


Fig. 3 Schematic diagram of the MAFLD patient evaluation model based on RF

powerful for image-based tasks, suffered from overfitting due to the limited dataset size and the complexity of end-to-end training, resulting in unstable test accuracy. Given its balanced accuracy, resistance to overfitting, and interpretability, the RF model was ultimately selected as the optimal classifier for liver fibrosis classification in this study.

Dataset splitting

To ensure the scientific rigor and fairness of model training and evaluation, the dataset was partitioned into training and testing subsets. Careful attention was paid to maintaining class balance and preventing data leakage during the splitting process. A stratified sampling strategy was employed to ensure an even distribution of samples representing different levels of fibrosis and disease severity across both subsets, thereby improving the model's generalizability and stability. The final train-test split ratio was set at 70:30, resulting in 218 samples in the training set and 94 samples in the testing set, after excluding outlier data. Furthermore, the proportions of low fibrosis and moderate-to-severe fibrosis cases were maintained at a 1:1 ratio in both subsets. This class balance is critical for mitigating classification bias and optimizing model performance. The training set was used to build and fine-tune the predictive model, while the testing set served as an independent evaluation cohort to assess final model performance on unseen data.

Model training

Model training was performed using the stratified training set. During this process, the cross-entropy loss function was selected to quantify the discrepancy between predicted probabilities and actual class labels. The Adam optimizer was adopted to accelerate convergence by adaptively adjusting the learning rate based on gradient estimates. To ensure robustness, we employed tenfold cross-validation during training, using cross-entropy as the evaluation metric. Key hyperparameters, such as the learning rate and the number of estimators, were fine-tuned using grid search. Throughout training, model performance was continuously monitored on a validation subset, and an early stopping strategy was implemented to prevent overfitting. Training was halted when no further improvement was observed on the validation set.

The final model performance was evaluated on the independent test set, which was completely isolated from the training and cross-validation processes. This approach provides an unbiased estimate of the model's generalization ability on previously unseen data. The RF classifier was implemented using the RandomForestClassifier class from the scikit-learn package (version 1.2.2) in Python 3.9.13. Hyperparameter tuning was conducted using GridSearchCV. All data preprocessing and model construction were performed in the Python environment, utilizing the following libraries: NumPy 1.23, Pandas 1.5, Matplotlib 3.7, and OpenCV 4.7.

Model validation and evaluation

Upon completion of model training, its classification performance was evaluated on the independent test set. Evaluation metrics included accuracy, sensitivity (recall), specificity, as well as the receiver operating characteristic (ROC) curve and AUC. Accuracy reflects the overall correctness of the model's predictions. Sensitivity measures the model's ability to correctly identify positive cases. Specificity indicates its ability to correctly classify negative cases. ROC-AUC provides a comprehensive measure of performance across varying classification thresholds. These metrics collectively offer a robust assessment of the model's predictive effectiveness and generalizability. In addition, for misclassified cases, error analysis was performed to identify potential model limitations, which in turn informed targeted optimization strategies to enhance performance and robustness.

Statistical analysis

All data preprocessing, analysis, and model development were conducted using Python in a Jupyter Notebook environment, with visualizations produced via Microsoft Visio.

The following Python libraries and statistical methods were employed: (1) Pandas: used for efficient data management, cleaning, and organization of MRI image paths and associated feature datasets. (2) NumPy: applied for data normalization and image processing tasks, such as histogram equalization, which enhances contrast and standardizes feature scales to stabilize model training. (3) Matplotlib: utilized for graphical representation of data distributions, model performance metrics, and feature maps.

To assess classification performance, we employed the following metrics: Confusion matrix: used to compute accuracy, precision, recall, and F1-score. ROC curves and AUC: used to evaluate the classifier's performance across various thresholds, offering a holistic view of model discrimination capability. These statistical tools and evaluation methods ensured a comprehensive, scientifically rigorous validation of the RF model's performance in classifying fibrosis severity among MAFLD patients.

Results

Data augmentation enhances the diversity of the MRI dataset and improves model training robustness

We collected MRI imaging data of MAFLD patients from the TCIA and Liver Imaging Database, comprising a total of 26 cases. Each case included four image sequences: T1-weighted, T2-weighted, in-phase, and opposed-phase images. All images were standardized and resized to a consistent resolution of 256 × 256 pixels.

To improve the diversity of the dataset and reduce the risk of model overfitting, we applied a series of data augmentation techniques, including image rotation, flipping, histogram equalization, gamma correction, and logarithmic transformation. After augmentation, the dataset size increased to 104 cases. These preprocessing steps effectively enhanced model robustness by introducing variations that simulate real-world imaging heterogeneity. Representative examples of images following histogram equalization, gamma correction, and logarithmic transformation are shown in Fig. 4.

Note: It is evident from Fig. 4 that each data enhancement results in different changes, significantly increasing our sample size and ensuring the model's performance.

Feature selection and dimensionality reduction optimization for improving the accuracy of MRI-based image analysis

We employed texture analysis techniques, including GLCM and texture entropy, to extract five key radiomics features—contrast, correlation, homogeneity, energy, and entropy—from each of the four MRI sequences. In total, 20 radiomics features were extracted per image. From the 26 patient cases, we obtained 518 image samples. After data cleaning, which excluded 216 low-quality or incomplete samples, 312 high-quality samples remained and were included in the analysis.

To optimize the model's performance and reduce dimensionality, we applied RFE for feature selection. This algorithm ranked features based on model-derived importance metrics (e.g., weight coefficients or Gini importance) and recursively eliminated the least relevant ones. The process continued until the top 10 most informative features were retained, which substantially reduced model complexity while preserving critical information. This dimensionality reduction approach provided several benefits: Enhanced model performance: Retaining high-value features improved generalizability and predictive accuracy. Improved interpretability: Reducing feature count simplified the model structure, making it more transparent and easier to interpret. Figure 5 illustrates the distribution of sample values for the ten selected features. The plots demonstrate clear intergroup differences and substantial feature variability, supporting their utility in subsequent classification modeling. Figure 6 displays a correlation heatmap of the selected features. As shown, the inter-feature correlations were generally low, with several negative correlations observed. This low collinearity further supports the appropriateness of the selected features for model construction.

Optimization of the RF model and hyperparameter tuning improve predictive accuracy

The RF algorithm was selected to construct the classification model. To optimize performance, a combination of grid search and tenfold cross-validation was used to fine-tune key hyperparameters. Incremental parameter adjustment identified the optimal configuration: the number of trees (n_estimators) was set to 43, and the maximum tree depth (max_depth) to 6, significantly enhancing both the model's generalization capacity and predictive accuracy. The hyperparameter tuning process is illustrated in Fig. 7.

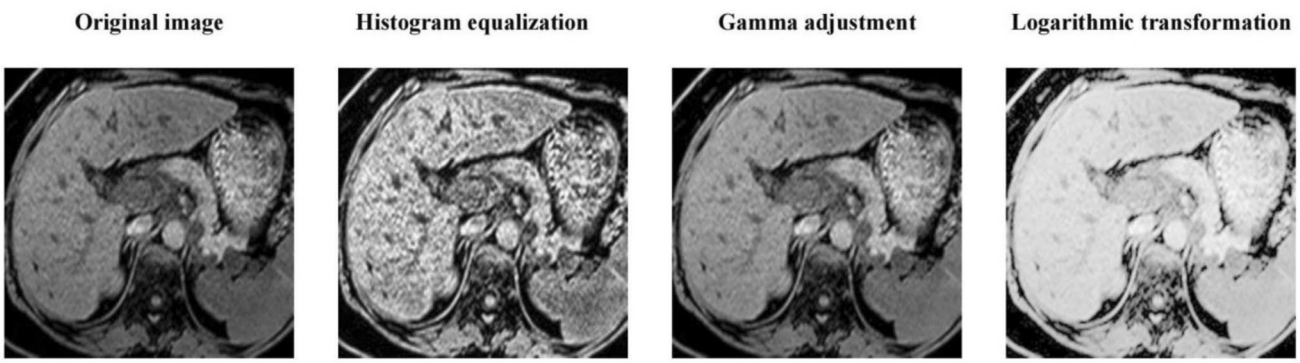


Fig. 4 Enhancement display of imaging data

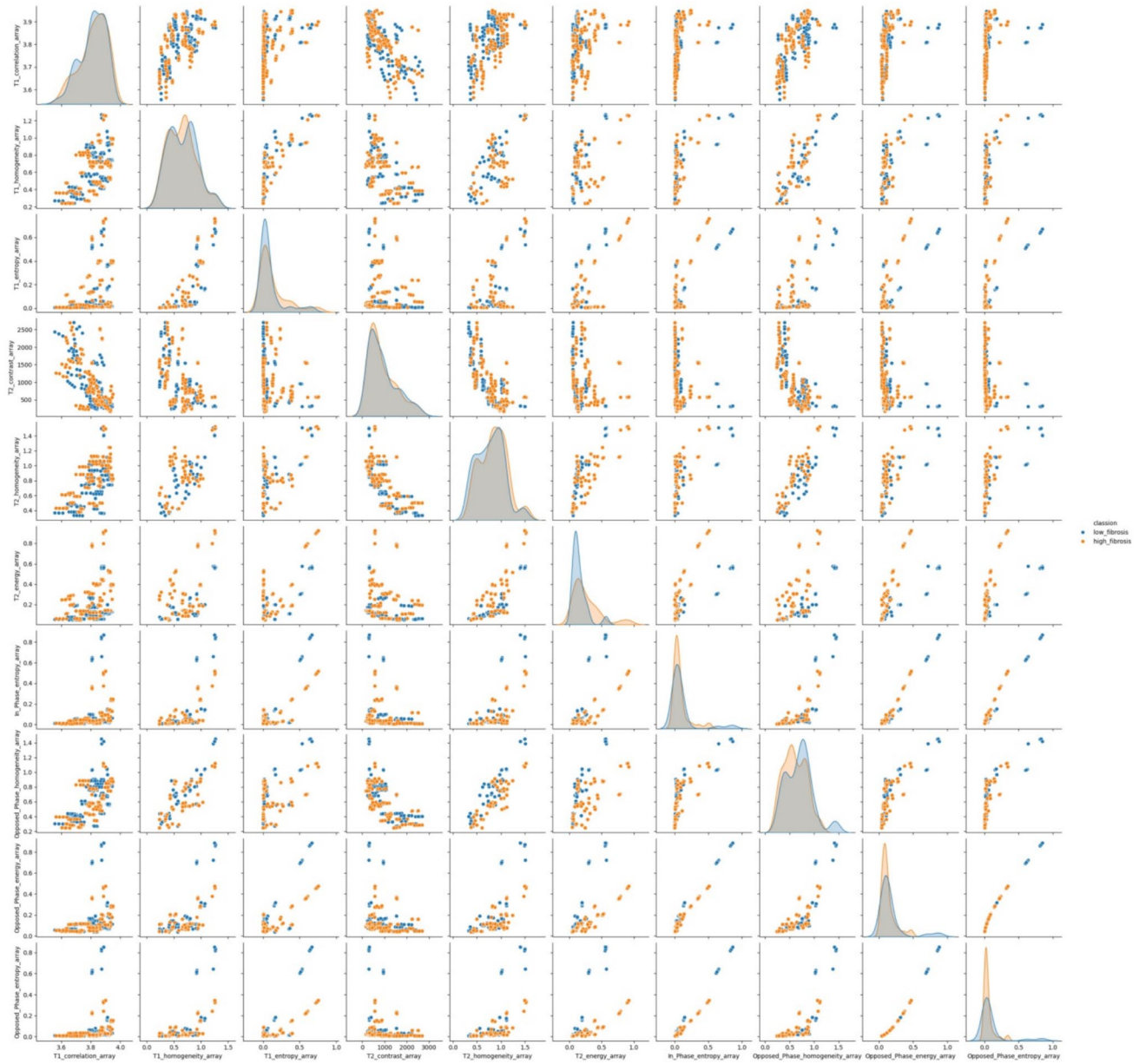


Fig. 5 Distribution plot of feature values

Model training was conducted using the training set, with cross-entropy as the loss function and the Adam optimizer to accelerate convergence. The training loss decreased steadily, indicating effective learning. An early stopping strategy was implemented by monitoring performance on the validation set in real time, which helped prevent overfitting and maintained stable accuracy across epochs.

Note: From the above figure, the optimal range for n_estimators is between 20 and 60. Combining with max_depth, the best parameter combination is determined as n_estimators = 43 and max_depth = 6.

High-accuracy RF model demonstrates excellent performance in classifying fibrosis severity in MAFLD

The trained RF model was evaluated on an independent test set to assess its ability to classify liver fibrosis severity in MAFLD patients (i.e., mild vs. moderate-to-severe fibrosis). The model achieved an accuracy of 96.8%, sensitivity of 95.7%, specificity of 97.8%, and an F1-score of 96.8%, reflecting its excellent discriminatory power. The confusion matrix visualizing classification performance is presented in Fig. 8.

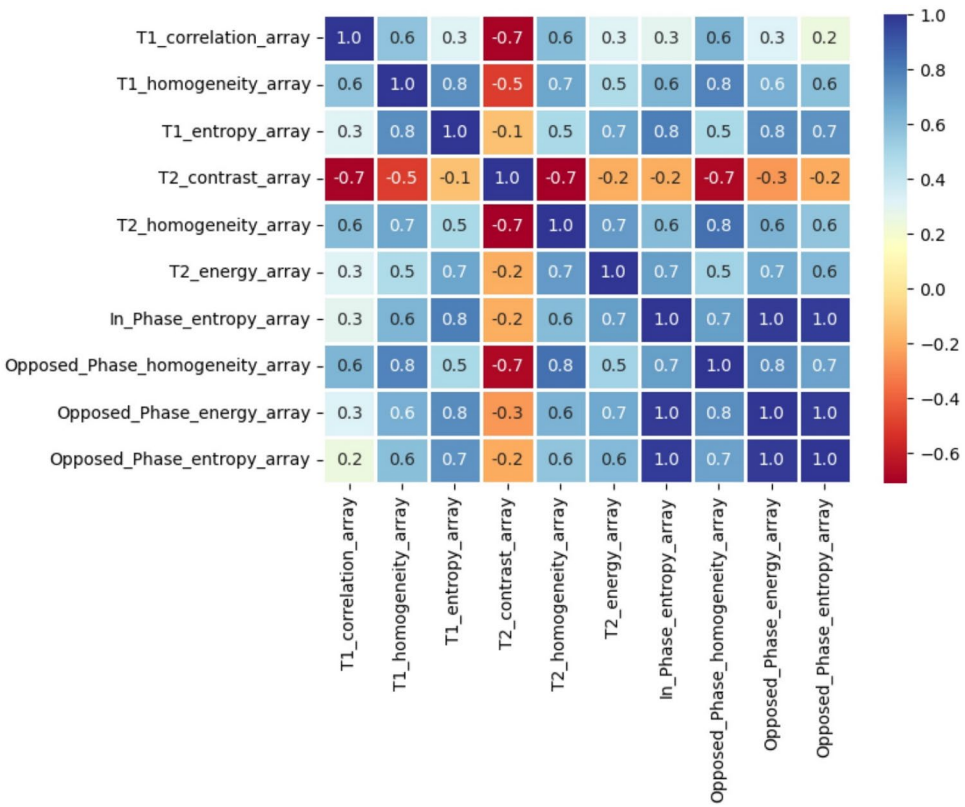


Fig. 6 Heatmap of feature correlation

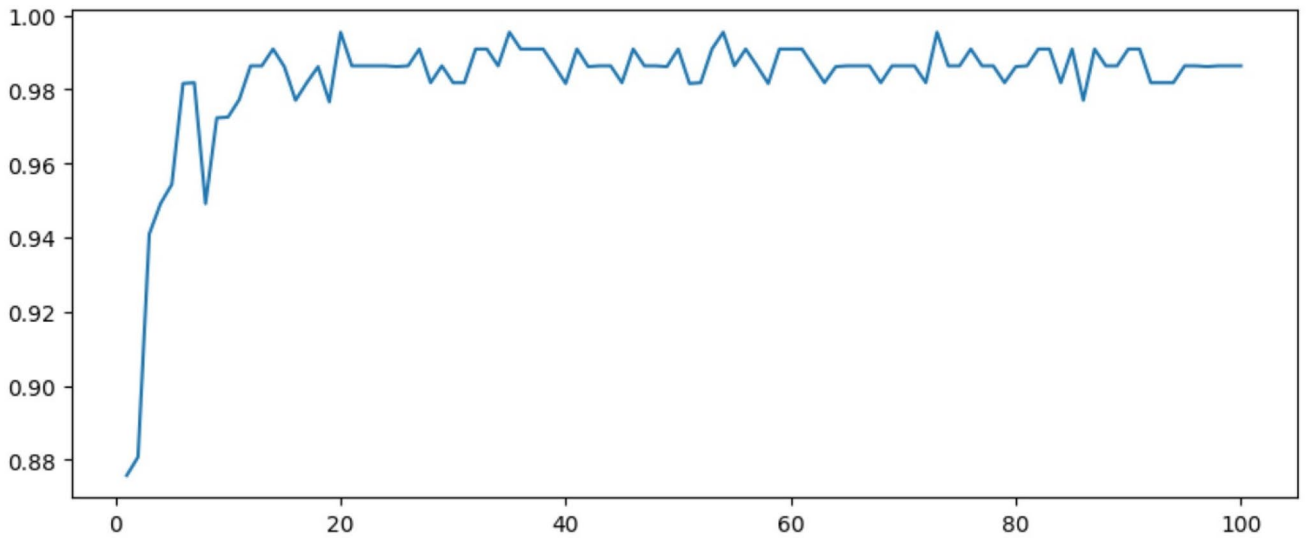


Fig. 7 Selection of the model hyperparameter n_estimators

Note: As shown in Fig. 8, the model made only a few misclassifications between low- and high-fibrosis cases (three instances in total): two high-fibrosis cases were predicted as low fibrosis, and one low-fibrosis case was predicted as high fibrosis. Further analysis revealed that these misclassified samples had feature value distributions close to the decision boundary between the two classes, particularly for the texture entropy and homogeneity variables, where their values deviated significantly from the respective group medians. Additionally, some of these images exhibited relatively low signal-to-noise ratios or tissue blurring, which may have compromised feature extraction accuracy. This suggests that incorporating an image quality assessment mechanism or integrating

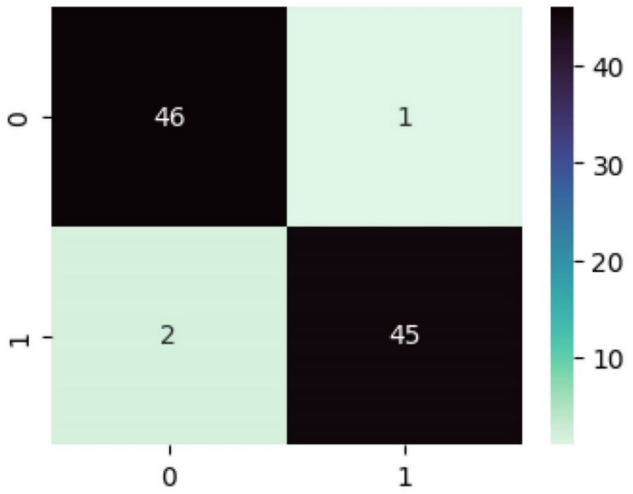


Fig. 8 Confusion matrix of model prediction results

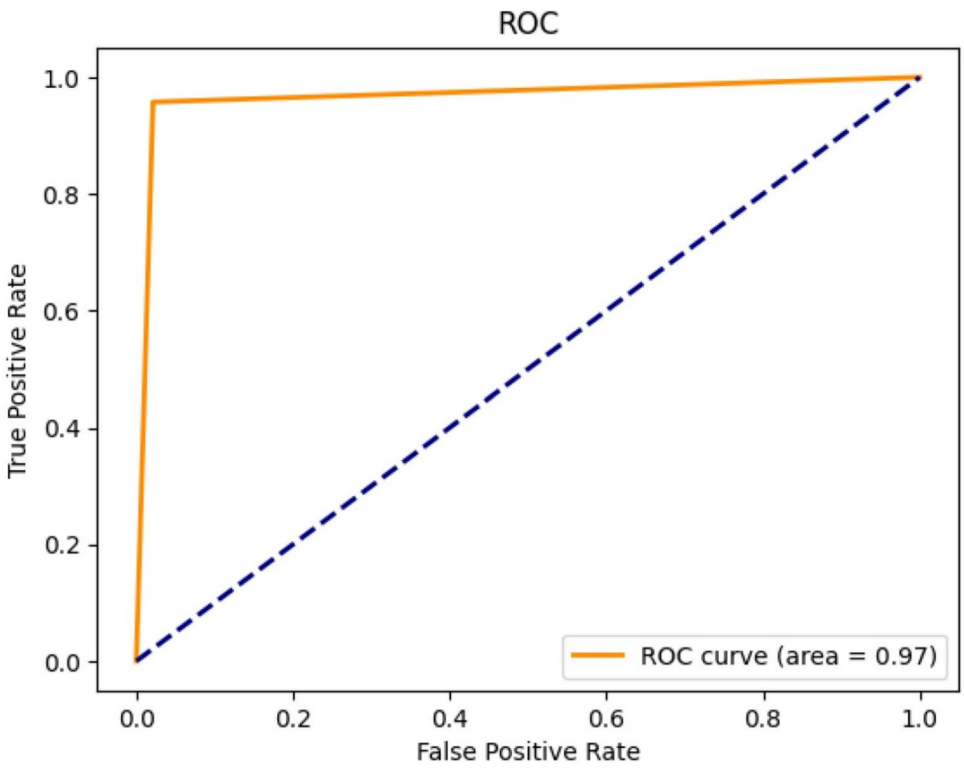
multi-timepoint dynamic MRI data may help improve future model performance.

Fibrosis-stage labels were based on FibroScan-derived liver stiffness measurements consistent with the F-staging system (F0-F4), as annotated in the source databases. The model's predicted outputs demonstrated strong concordance with FibroScan-based labels, achieving an AUC of 0.97 (Fig. 9). These findings indicate high potential for the RF model as a noninvasive diagnostic tool for assessing liver fibrosis in MAFLD patients. Importantly, all performance metrics were computed using the independent test set, which was entirely excluded from training and cross-validation. This approach ensured the objectivity and external validity of the reported evaluation results.

Fibrosis assessment in MAFLD patients

In this study, the importance of radiomic features was ranked and visualized based on their Gini importance scores derived from the RF model. The results showed that the energy feature extracted from the T2-weighted sequence, the homogeneity feature from T2, and the texture entropy feature from T1 had the highest weights in the model's decision-making process, underscoring their critical roles in distinguishing between different stages of fibrosis severity. Gini importance reflects the average information gain a feature contributes during node splitting across all decision trees in the ensemble. We further analyzed the ten key imaging features retained through the RFE method, which were subsequently used for model training (Fig. 10). Detailed descriptions of these features are provided in Supplementary Table 1. Among them, the T2_energy feature had the highest importance score, highlighting its pivotal role in fibrosis classification—likely due to the heightened sensitivity of T2-weighted imaging to changes in hepatic water content. Additionally, the T2_homogeneity and T1_entropy features also demonstrated strong discriminative power, suggesting

Fig. 9 ROC curve of the model



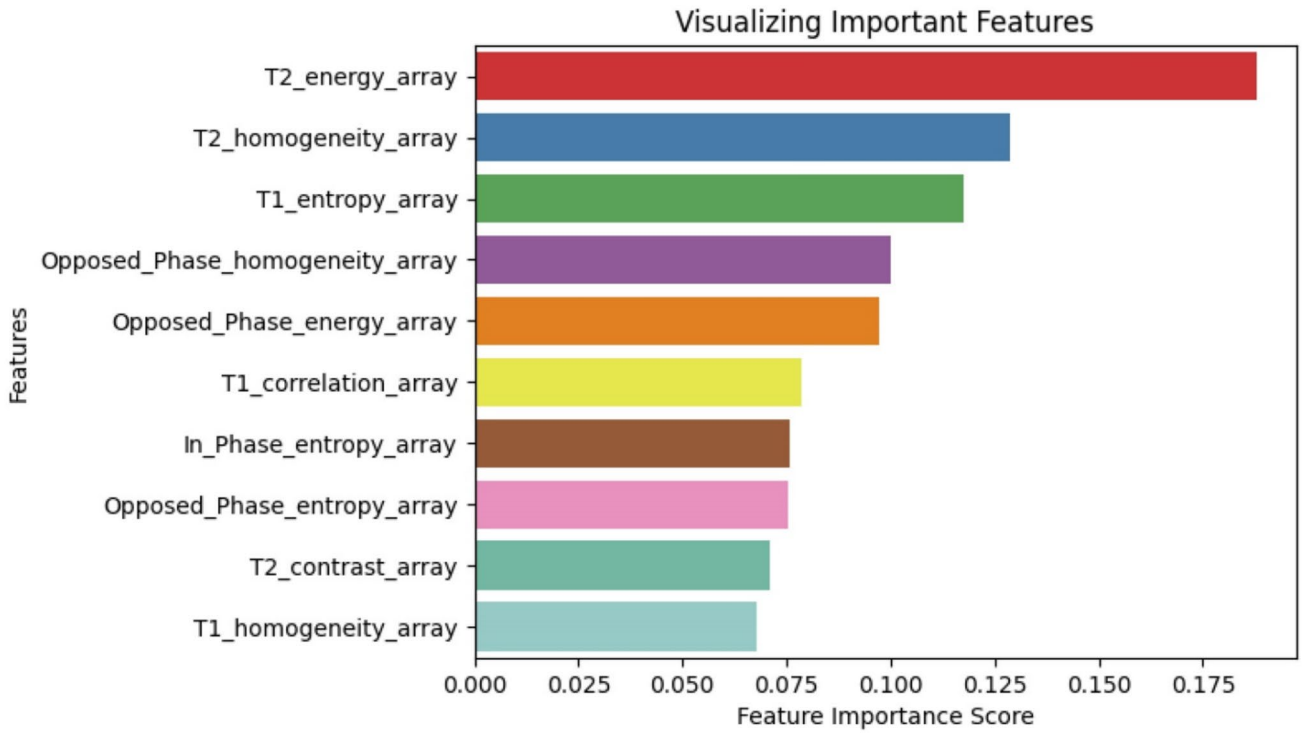


Fig. 10 Visualization of feature importance

that both tissue uniformity and microstructural complexity are relevant biomarkers in the progression of liver fibrosis. Features extracted from the opposed-phase modality, such as homogeneity and energy, also contributed substantially to the model's performance. These may reflect signal perturbations in fat–water interface regions, which are commonly altered in steatotic and fibrotic tissues.

Overall, these high-importance features captured multidimensional, multimodal textural alterations associated with MAFLD-related fibrosis, providing a biological rationale for the model's predictive mechanism.

To validate the predictive accuracy and stability of the model based on these features, a regression analysis was performed. The results indicated that the model's prediction errors remained within acceptable limits. Furthermore, comparative analysis between the model outputs and clinical assessments of fibrosis severity revealed a high degree of correlation (Fig. 11), supporting the model's reliability in real-world diagnostic settings. These findings affirm the robust performance of the RF-based evaluation model for MAFLD, particularly in assessing fibrosis and disease severity.

Early identification of fibrosis is critical for improving outcomes. Our results suggest that the model can significantly enhance early diagnosis rates of MAFLD, enabling timely clinical intervention. Moreover, by facilitating precise disease stratification and supporting evidence-based decision-making, this model offers substantial utility in optimizing treatment strategies and improving long-term prognosis in MAFLD patients.

Note: (A) Correlation analysis between energy feature information extracted from the T2 imaging mode and homogeneity feature information extracted from the T2 imaging mode with the model's predicted results and the actual severity of MAFLD fibrosis; (B) Correlation analysis between energy feature information extracted from the T2 imaging mode and texture entropy feature information extracted from the T1 imaging mode with the model's predicted results and the actual severity of MAFLD fibrosis; (C) Correlation analysis between homogeneity feature information extracted from the T2 imaging mode and texture entropy feature information extracted from the T2 imaging mode with the model's predicted results and the actual severity of MAFLD fibrosis.

Discussion

The RF model constructed in this study demonstrated high performance in evaluating liver fibrosis and disease severity in MAFLD patients, achieving an accuracy of 96.8%, sensitivity of 95.7%, specificity of 97.8%, and an F1-score of 96.8%. Compared with previous studies that employed traditional radiomics techniques, our approach significantly

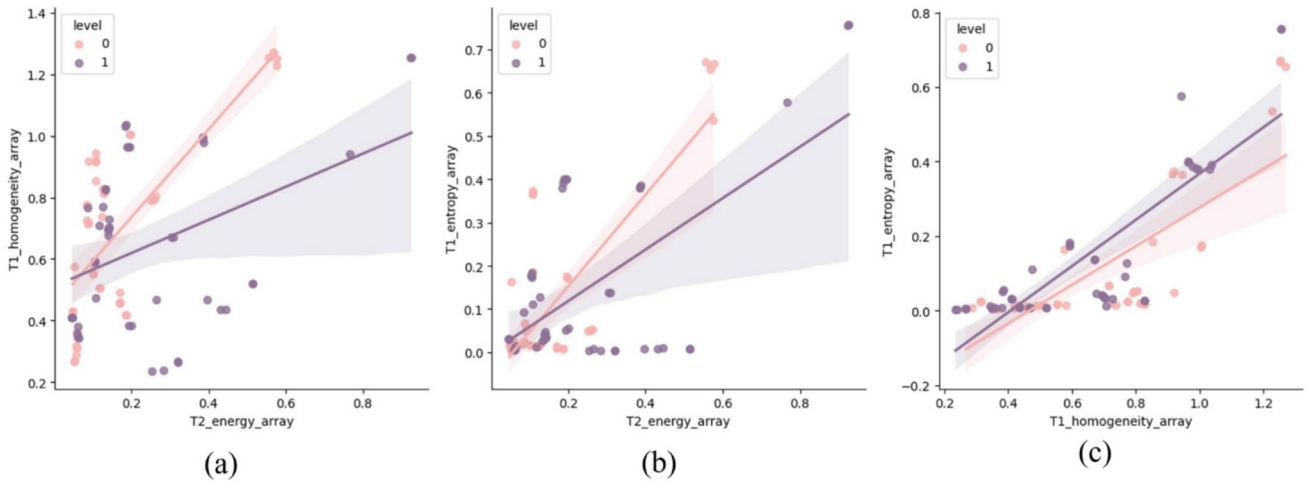


Fig. 11 Regression validation based on three key features

improved both the efficiency and accuracy of MRI image analysis by integrating advanced ML algorithms. Earlier methods often relied on manual feature extraction and basic statistical analyses, which were limited by subjective feature selection and analytical bias [38–40]. By employing automated feature extraction and systematic selection strategies, this study addressed those limitations and provided a more objective and reproducible framework for MAFLD diagnosis.

Specifically, we utilized GLCM techniques to extract five key texture features—contrast, correlation, homogeneity, energy, and entropy—from multimodal MRI sequences. These features provided a more comprehensive representation of liver tissue texture characteristics than traditional methods, enhancing the diagnostic resolution of imaging data. Additionally, RFE was applied for dimensionality reduction, significantly improving model efficiency while preserving the most predictive variables. Unlike prior studies where such systematic selection methods were often underutilized, our study incorporated RFE to streamline model complexity and strengthen interpretability [41–43].

The RF model, a robust ensemble learning method, was chosen as the primary classifier due to its proven efficacy in handling high-dimensional, heterogeneous imaging datasets [44, 45]. Compared to other ML techniques such as support vector machines (SVMs) or deep neural networks (DNNs), RF models offered superior interpretability, reduced susceptibility to overfitting, and effective handling of imbalanced data distributions [46–48]. Furthermore, RF models are computationally efficient, require less parameter tuning, and do not necessitate large-scale annotated datasets or high-performance computing resources—making them practically advantageous in clinical settings. While the RF model exhibited excellent performance in this study, we acknowledged its limitations and planned to incorporate a broader range of algorithms—including XGBoost, CNNs, and transformers—in future investigations. This comparative strategy aims to systematically evaluate model generalizability and optimize algorithm selection for various clinical contexts. In contrast to prior research that typically focused on a single algorithm, our study undertook comparative assessments and ultimately identified the RF model as the optimal approach for MAFLD fibrosis classification. This not only provided a reliable non-invasive diagnostic tool, but also laid the groundwork for future AI-assisted clinical decision support systems in liver disease management.

This study conducted a comprehensive investigation into data preprocessing and model optimization, both of which are critical yet often underemphasized components in ML research. Initially, all MRI images were standardized in terms of resolution, contrast, and size to ensure consistency and comparability across datasets. Following this, a tenfold cross-validation approach was employed to fine-tune two key hyperparameters of the RF model—n_estimators and max_depth—to determine the optimal parameter configuration. In contrast to prior studies that frequently overlooked rigorous preprocessing and hyperparameter tuning, thereby compromising model robustness and predictive performance [49, 50], our methodical approach significantly enhanced the accuracy, reliability, and generalization ability of the model.

Furthermore, this study systematically compared the model's classification outcomes with histopathological findings, revealing a high degree of consistency between the two. This concordance underscores the clinical relevance and diagnostic utility of ML-based image analysis, supporting its role as a noninvasive adjunctive tool for evaluating fibrosis and disease severity in MAFLD patients. While prior studies have attempted to apply ML to liver imaging, they often lacked rigorous external validation or direct comparison with pathological standards, limiting their real-world

applicability [26, 43, 51]. Despite these promising results, one key limitation of the current work is the absence of external validation. The model was trained and validated exclusively on internal data, which may limit its generalizability to broader clinical populations. To address this, future research will incorporate external datasets derived from independent clinical centers or real-world patient cohorts. This strategy will enable more robust testing of the model's transferability, scalability, and clinical value in diverse healthcare settings.

This study demonstrated the potential of combining multimodal MRI imaging with ML algorithms for the accurate and noninvasive assessment of liver fibrosis in MAFLD patients. The approach not only improves diagnostic efficiency, but also provides a reproducible and interpretable framework for clinical decision support. The high accuracy and reliability achieved in this study pave the way for broader clinical adoption and lay the foundation for extending this model to other liver pathologies, such as viral hepatitis, alcoholic liver disease, and autoimmune liver conditions. Thus, the proposed framework holds significant translational potential across a wide spectrum of hepatic diseases.

This study successfully performed a quantitative analysis of MRI data from MAFLD patients by developing a RF ML model, effectively evaluating the relationship between hepatic fat content and the degree of liver fibrosis and disease severity. The model demonstrated outstanding classification performance—achieving high accuracy, sensitivity, and specificity—which highlights its strong potential as a noninvasive diagnostic tool in clinical practice. Scientifically, the study contributes novel insights into the application of ML in medical image analysis, advancing the integration of radiomics and artificial intelligence in hepatic disease evaluation. Clinically, the model offers an effective, noninvasive alternative to traditional pathological assessment, thereby supporting more accurate diagnosis and treatment decision-making while reducing patient burden and procedural risks.

In addition, preliminary error analysis revealed that misclassifications were primarily attributable to ambiguous texture patterns in borderline cases and variations in image quality. To address these issues, future studies should pursue two key directions: (1) refining image quality control and preprocessing pipelines, and (2) incorporating uncertainty estimation mechanisms or ensemble modeling strategies to better quantify classification confidence, particularly for borderline cases. Moreover, due to the absence of complete histopathological scoring (e.g., METAVIR or NAS criteria) in the dataset, this study relied on FibroScan-based liver stiffness measurements as surrogate labels. While clinically meaningful, this substitution may introduce bias. Future work should integrate standardized histopathological scoring systems to more rigorously validate the correspondence between model predictions and gold-standard histological findings.

Despite the promising results, several limitations merit discussion. First, as a retrospective study using publicly available datasets, there were inherent constraints in patient selection, image acquisition protocols, and clinical follow-up. Potential confounding variables and therapeutic interventions could not be fully accounted for, which may influence the generalizability of the findings. Second, the dataset originated from a relatively limited and homogeneous patient population in terms of ethnicity, imaging platforms, and scanning parameters. This restricts the model's applicability to more diverse real-world clinical settings. Thus, large-scale, prospective, multicenter studies incorporating multi-ethnic and heterogeneous patient cohorts are essential to confirm the model's robustness and broad clinical utility.

Although the model demonstrated excellent performance on an independent test set and exhibits strong potential for clinical deployment, several practical challenges remain. Variability in MRI acquisition across institutions—such as differences in scanner type, resolution, contrast enhancement protocols, and imaging parameters—may adversely affect feature extraction and model consistency. Additionally, real-world implementation requires significant computational resources, technical infrastructure, and specialized personnel for data preprocessing and deployment, which may pose barriers in resource-limited or community-based healthcare settings. Addressing these challenges will require the development of standardized imaging protocols and lightweight, scalable model deployment solutions to facilitate widespread adoption.

Looking ahead, future research should prioritize expanding the dataset and increasing its diversity by incorporating multicenter, multi-regional cohorts across different demographic backgrounds. Moreover, exploring advanced algorithms such as XGBoost, transformers, or hybrid deep learning architectures may further enhance predictive performance and robustness. Parallel efforts should aim to simplify data preprocessing workflows and reduce computational demands, making the system more accessible to primary care settings. These advancements may contribute to the development of personalized and precision-based diagnostic pathways for MAFLD, ultimately improving clinical outcomes and quality of life for affected patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10238-025-01818-5.

Author contributions Mengting Hou and Yujie Zhu contributed equally to data collection, image preprocessing, and radiomics feature extraction. Huadi Zhou and Siyi Zhou were responsible for model development, performance evaluation, and statistical analysis. Jianjun Zhang assisted in clinical data interpretation and validation. Yue Zhang and Xiao Liu supervised the project, provided critical revisions, and

corresponded with the journal. All authors reviewed and approved the final manuscript.

Funding The authors have not disclosed any funding.

Data availability Data is provided within the manuscript or supplementary information files.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval and consent to participate This study exclusively used MRI imaging data obtained from public databases (TCIA and Liver Imaging Database). No prospective patient enrollment or clinical intervention was involved; therefore, ethical approval was not required.

References

1. Guo X, Yin X, Liu Z, Wang J. Non-alcoholic fatty liver disease (NAFLD) pathogenesis and natural products for prevention and treatment. Int J Mol Sci. 2022;23:15489.
2. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, et al. A multisociety delphi consensus statement on new fatty liver disease nomenclature. Hepatology. 2023;78:1966–86.
3. Yuan S, Chen J, Ruan X, Sun Y, Zhang K, Wang X, Larsson SC. Smoking, alcohol consumption, and 24 gastrointestinal diseases: mendelian randomization analysis. Elife. 2023;12:e84051.
4. Juanola O, Martínez-López S, Francés R, Gómez-Hurtado I. Non-alcoholic fatty liver disease: metabolic, genetic, epigenetic and environmental risk factors. Int J Environ Res Public Health. 2021;18:5227.
5. Muzurović EM, Volčanšek Š, Tomšić KZ, Janež A, Mikhailidis DP, Rizzo M, et al. Glucagon-like peptide-1 receptor agonists and dual glucose-dependent insulinotropic polypeptide/glucagon-like peptide-1 receptor agonists in the treatment of obesity/metabolic syndrome, prediabetes/diabetes and non-alcoholic fatty liver disease—current evidence. J Cardiovasc Pharm Therapeutics. 2022;27:10742484221146372.
6. Torres-Peña JD, Arenas-de Larriva AP, Alcala-Diaz JF, Lopez-Miranda J, Delgado-Lista J. Different dietary approaches, non-alcoholic fatty liver disease and cardiovascular disease: a literature review. Nutrients. 2023;15:1483.
7. Powell EE, Wong VW-S, Rinella M. Non-alcoholic fatty liver disease. Lancet. 2021;397:2212–24.
8. Paternostro R, Trauner M. Current treatment of non-alcoholic fatty liver disease. J Intern Med. 2022;292:190–204.
9. Tincopa MA, Loomba R. Non-invasive diagnosis and monitoring of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis. Lancet Gastroenterol Hepatol. 2023;8:660–70.
10. Loomba R, Abdelmalek MF, Armstrong MJ, Jara M, Kjær MS, Krarup N, et al. Semaglutide 2·4 mg once weekly in patients with non-alcoholic steatohepatitis-related cirrhosis: a randomised, placebo-controlled phase 2 trial. Lancet Gastroenterol Hepatol. 2023;8:511–22.
11. Sanyal AJ, Van Natta ML, Clark J, Neuschwander-Tetri BA, Diehl A, Dasarathy S, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. N Engl J Med. 2021;385:1559–69.
12. Younossi ZM, Paik JM, Al Shabeeb R, Golabi P, Younossi I, Henry L. Are there outcome differences between NAFLD and metabolic-associated fatty liver disease? Hepatology. 2022;76:1423–37.
13. Wu X, Cheung CKY, Ye D, Chakrabarti S, Mahajan H, Yan S, et al. Serum thrombospondin-2 levels are closely associated with the severity of metabolic syndrome and metabolic associated fatty liver disease. J Clin Endocrinol Metab. 2022;107: e3230–40.
14. Xie R, Xiao M, Li L, Ma N, Liu M, Huang X, et al. Association between SII and hepatic steatosis and liver fibrosis: a population-based study. Front Immunol. 2022. https://doi.org/10.3389/fimmu.2022.925690.
15. Govaere O, Hasoon M, Alexander L, Cockell S, Tiniakos D, Ekstedt M, et al. A proteo-transcriptomic map of non-alcoholic fatty liver disease signatures. Nat Metab. 2023;5:572–8.
16. Ozturk A, Olson MC, Samir AE, Venkatesh SK. Liver fibrosis assessment: MR and US elastography. Abdom Radiol. 2021;47:3037–50.
17. Ajmera V, Loomba R. Imaging biomarkers of NAFLD, NASH, and fibrosis. Mol Metab. 2021;50: 101167.
18. Starekova J, Hernando D, Pickhardt PJ, Reeder SB. Quantification of liver fat content with CT and MRI: state of the art. Radiology. 2021;301:250–62.
19. Anani T, Rahmati S, Sultana N, David AE. MRI-traceable theranostic nanoparticles for targeted cancer treatment. Theranostics. 2021;11:579–601.
20. Johnston EW, Fotiadis N, Cummings C, Basso J, Tyne T, Lameijer J, et al. Developing and testing a robotic MRI/CT fusion biopsy technique using a purpose-built interventional phantom. Euro Radiol Exper. 2022;6:55.
21. Liang X, Fu Y, Cao W, Wang Z, Zhang K, Jiang Z, et al. Gut microbiome, cognitive function and brain structure: a multi-omics integration analysis. Transl Neurodegener. 2022. https://doi.org/10.1186/s40035-022-00323-z.
22. Yang Y, Knol MJ, Wang R, Mishra A, Liu D, Luciano M, et al. Epigenetic and integrative cross-omics analyses of cerebral white matter hyperintensities on MRI. Brain. 2022;146:492–506.
23. Gajjar A, Robinson GW, Smith KS, Lin T, Merchant TE, Chintagumpala M, et al. Outcomes by clinical and molecular features in children with medulloblastoma treated with risk-adapted therapy: results of an international phase III trial (SJMB03). J Clin Oncol. 2021;39:822–35.
24. Dong S, Luo G, Tam C, Wang W, Wang K, Cao S, et al. Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. Med Image Anal. 2020;61: 101638.
25. Zeinali N, Youn N, Albashayreh A, Fan W, Gilbertson White S. Machine learning approaches to predict symptoms in people with cancer: systematic review. JMIR Cancer. 2024;10: e52322.
26. Chen X, Wang X, Zhang K, Fung K-M, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal. 2022;79: 102444.
27. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E, et al. Deep learning for caries detection: a systematic review. J Dent. 2022;122: 104115.

28. Zheng X, He B, Hu Y, Ren M, Chen Z, Zhang Z, et al. Diagnostic accuracy of deep learning and radiomics in lung cancer staging: a systematic review and meta-analysis. Front Public Health. 2022. https://doi.org/10.3389/fpubh.2022.938113.
29. Boudreault J, Campagna C, Chebana F. Machine and deep learning for modelling heat-health relationships. Sci Total Environ. 2023;892: 164660.
30. Yu W, Li S, Ye T, Xu R, Song J, Guo Y. Deep ensemble machine learning framework for the estimation of PM2.5 concentrations. Environ Health Perspect. 2022. https://doi.org/10.1289/EHP9752.
31. Kwon Y, Lee J, Park JH, Kim YM, Kim SH, Won YJ, et al. Osteoporosis pre-screening using ensemble machine learning in postmenopausal Korean women. Healthcare. 2022;10:1107.
32. Takahashi S, Terai H, Hoshino M, Tsujio T, Kato M, Toyoda H, et al. Machine-learning-based approach for nonunion prediction following osteoporotic vertebral fractures. Eur Spine J. 2022;32:3788–96.
33. Mlambo F, Chironda C, George J. Risk stratification of COVID-19 using routine laboratory tests: a machine learning approach. Infect Dis Rep. 2022;14:900–31.
34. Lu X, Du J, Zheng L, Wang G, Li X, Sun L, et al. Feature fusion improves performance and interpretability of machine learning models in identifying soil pollution of potentially contaminated sites. Ecotoxicol Environ Saf. 2023;259: 115052.
35. Nguyen MB, Dragulescu A, Chaturvedi R, Fan C-PS, Villemain O, Friedberg MK, et al. Understanding complex interactions in pediatric diastolic function assessment. J Am Soc Echocardiogr. 2022;35:868-877.e5.
36. Chang D, Truong E, Mena EA, Pacheco F, Wong M, Guindi M, et al. Machine learning models are superior to noninvasive tests in identifying clinically significant stages of NAFLD and NAFLD-related cirrhosis. Hepatology. 2022;77:546–57.
37. Bettini S, Serra R, Fabris R, Dal Prà C, Favaretto F, Dassie F, et al. Association of obstructive sleep apnea with non-alcoholic fatty liver disease in patients with obesity: an observational study. Eat Weight Disord-Stud Anorex Bulim Obes. 2021;27:335–43.
38. Alibabaei S, Rahmani M, Tahmasbi M, Tahmasebi Birgani MJ, Razmjoo S. Evaluating the gray level co-occurrence matrix-based texture features of magnetic resonance images for glioblastoma multiform patients' treatment response assessment. J Med Signals Sens. 2023;13:261–71.
39. Obaloluwa Olaniyi E. Eye melanoma diagnosis system using statistical texture feature extraction and soft computing techniques. In: Journal of Biomedical Physics and Engineering. 2022
40. Athertya JS, Saravana Kumar G. Classification of certain vertebral degenerations using MRI image features. Biomed Phys Eng Express. 2021;7: 045013.
41. Scapicchio C, Gabelloni M, Barucci A, Cioni D, Saba L, Neri E. A deep look into radiomics. Radiol Med. 2021;126:1296–311.
42. Wu G, Jochems A, Refaee T, Ibrahim A, Yan C, Sanduleanu S, et al. Structural and functional radiomics for lung cancer. Eur J Nucl Med Mol Imaging. 2021;48:3961–74.
43. Dar RA, Rasool M, Assad A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Comput Biol Med. 2022;149:106073.
44. Demir B, Ulukaya S, Erdem O. Detection of Parkinson's disease with keystroke data. Comput Methods Biomech Biomed Eng. 2023;26:1653–67.
45. Jose DM, Vincent AM, Dwarakish GS. Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. Sci Rep. 2022. https://doi.org/10.1038/s41598-022-08786-w.
46. Li W, Guan X, Wang Y, Lv Y, Wu Y, Yu M, et al. Cuproptosis-related gene identification and immune infiltration analysis in systemic lupus erythematosus. Front Immunol. 2023. https://doi.org/10.3389/fimmu.2023.1157196.
47. Wei T-T, Zhang J-F, Cheng Z, Jiang L, Li J-Y, Zhou L. Development and validation of a machine learning model for differential diagnosis of malignant pleural effusion using routine laboratory data. Therapeutic Adv Respiratory Dis. 2023;17:17534666231208632.
48. 基于DNA甲基化推断年龄的建模方法与影响因素. 法医学杂志. 2023;39.
49. Stanosheck JA, Castell-Perez ME, Moreira RG, King MD, Castillo A. Oversampling methods for machine learning model data training to improve model capabilities to predict the presence of *Escherichia coli* MG1655 in spinach wash water. J Food Sci. 2023;89:150–73.
50. Bezjak M, Kocman B, Jadrijević S, Kanižaj TF, Antonijević M, Dalbelo Bašić B, et al. Use of machine learning models for identification of predictors of survival and tumour recurrence in liver transplant recipients with hepatocellular carcinoma. Ann Transl Med. 2023;11:345–345.
51. Yousif M, van Diest PJ, Laurinavicius A, Rimm D, van der Laak J, Madabhushi A, et al. Artificial intelligence applied to breast pathology. Virchows Arch. 2021;480:191–209.