Endocrine (2024) 85:615–625
https://doi.org/10.1007/s12020-024-03735-1

ORIGINAL ARTICLE

# Integrated machine learning and deep learning for predicting diabetic nephropathy model construction, validation, and interpretability

Junjie Ma[1] · Shaoguang An[1] · Mohan Cao[1] · Lei Zhang[2] · Jin Lu[3,4]

Received: 13 November 2023 / Accepted: 6 February 2024 / Published online: 23 February 2024


## Abstract

**Objective** To construct a risk prediction model for assisted diagnosis of Diabetic Nephropathy (DN) using machine learning algorithms, and to validate it internally and externally.

**Methods** Firstly, the data was cleaned and enhanced, and was divided into training and test sets according to the 7:3 ratio. Then, the metrics related to DN were filtered by difference analysis, Least Absolute Shrinkage and Selection Operator (LASSO), Recursive Feature Elimination (RFE), and Max-relevance and Min-redundancy (MRMR) algorithms. Ten machine learning models were constructed based on the key variables. The best model was filtered by Receiver Operating Characteristic (ROC), Precision-Recall (PR), Accuracy, Matthews Correlation Coefficient (MCC), and Kappa, and was internally and externally validated. Based on the best model, an online platform had been constructed.

**Results** 15 key variables were selected, and among the 10 machine learning models, the Random Forest model achieved the best predictive performance. In the test set, the area under the ROC curve was 0.912, and in two external validation cohorts, the area under the ROC curve was 0.828 and 0.863, indicating excellent predictive and generalization abilities.

**Conclusion** The model has a good predictive value and is expected to help in the early diagnosis and screening of clinical DN.

**Keywords** Diabetic nephropathy · Machine learning · Clinical prediction model · Interpretability

## Introduction

Diabetic Nephropathy (DN) refers to a significant complication characterized by structural and functional changes in the kidneys induced by diabetes. DN has emerged as the primary driving factor behind renal failure, with approximately 40% of cases of renal failure attributed to DN [1]. As of 2017, the global prevalence of DN was estimated to be approximately 9.1%, with around 1.2 million patients succumbing to DN-related complications [2]. Research demonstrated that targeted interventions during the early stages of DN effectively prevented or slowed the progression of renal failure and improved patient outcomes [3]. However, the occurrence and development of DN are influenced by a variety of complex pathophysiological mechanisms, including metabolic disturbances, genetic factors, inflammation, and oxidative stress, which leads to challenges in clinical diagnosis and treatment [4, 5]. Currently, the clinical diagnosis of DN primarily relies on serum creatinine levels and urinary albumin values. However, the accuracy of these measurements may be compromised due to biological limitations and analytical variability [6]. Therefore, there is an urgent need to develop and validate novel diagnostic approaches for DN to achieve early detection, diagnosis, and treatment, which holds significant clinical relevance.

Machine Learning (ML), a multidisciplinary cross-technology, processes real-world data using various computer algorithms by simulating human learning behavior [7]. In

✉ Jin Lu
0100197@bbmc.edu.cn

1 Department of Clinical Medicine, Bengbu Medical University, Bengbu, China

2 Department of Oncology Surgery, the Second Affiliated Hospital of Bengbu Medical University, Bengbu, China

3 Anhui Key Laboratory of Computational Medicine and Intelligent Health, Bengbu Medical University, Bengbu, China

4 School of Basic Medicine, Bengbu Medical University, Bengbu, China

traditional healthcare practices, factors such as inadequate disease prevention and the inherent bias in a physician's experience can often lead to misdiagnosis, subsequently resulting in treatment failures or exacerbation of the medical condition [8]. With its high accuracy and big data processing capability, machine learning is conducive to alleviating the shortage of healthcare resources and the burden on personnel. Presently, the application of machine learning algorithmic models in the medical field is gradually increasing, including the early prevention and diagnosis of diseases [9], drug development [10], and infectious disease prevention and treatment [11], which show good prediction ability.

Currently, there are various studies applying machine learning models to the diagnosis of DN. Such as Yin et al. [12] used the XGboost model to predict the occurrence of DN based on serum metabolite levels. XU et al. [13] screened three DN-associated immune and oxidative stress genes based on bioinformatics and machine learning methods, which are expected to be pivotal genes for DN treatment. LIU et al. [14] comprehensively compared 15 machine learning models and finally incorporated the Catboost model for risk prediction of DN. Furthermore, Hosseini et al. [15] developed and validated a risk prediction model for type 2 diabetic nephropathy through a logistic regression model and developed an online tool to predict the risk score of diabetic nephropathy. Other scholars recognized fundus photographs of diabetic retinopathy by deep learning methods, which were used to predict diabetic end-stage renal disease, and the model performed well [16]. Based on their studies, it is significant to apply machine learning and deep learning models to the clinical diagnosis and screening of diabetic nephropathy. However, no comparative study comparing machine learning models with deep learning models in DN prediction has been observed. So, we synthesized machine learning and deep learning models for DN risk prediction, incorporated a wider variety of models to better explore the models suitable for our study, and validated the models as well as analyzed them for interpretability. In addition, we developed an online prediction platform based on the best models to be used by other researchers.

## Materials and methods

### Data source

#### Training data

The data is obtained from the National Population Health Data Center (NPHDC) of China (https://www.ncmi.cn). This dataset comprises 87 clinical variables for 3000 patients with type 2 diabetes, including general demographic information, physical examination data, laboratory data, and diabetes-related complications data. Within this dataset, there are 1277 samples of DN (Diabetic Nephropathy) and 1723 samples of non-DN (nDN).

#### External validation data

Data from the National Health and Nutrition Examination Survey (NHANES) of the United States (https://www.cdc.gov/nchs/nhanes) were included by the following criteria: (1) Downloading demographic data, physical examination data, laboratory data, and questionnaire data from 2015 to 2020; (2) Retaining only samples with a diagnosis of diabetes. In total, data from 1981 diabetic patients were obtained.

The Taiwan Biobank (TWBB) is a large-scale longitudinal study based on medical centers and several local chronic disease patient cohorts. It comprises extensive genomic and clinical examination data [17]. Clinical data from 3183 patients with type 2 diabetes were obtained through a database application for external validation.

### Data cleaning

The data used in the study were obtained from the clinical data of real diabetic patients with some missing values and outliers. Therefore, to ensure the reliability and standardization of the results, the following processing of the raw data is required: (1) Outliers: outliers were identified by the Interquartile Range (IQR) method. The upper bound of the distribution of the data is set as $Q3 + 1.5*(Q3–Q1)$ and the lower bound as $Q1–1.5*(Q3–Q1)$. For identified outliers in the data, we replace them with boundary values; (2) Missing values: variables and samples with more than 50% missing values are deleted, and the rest of the missing values are filled in by multiple regression interpolation with the "mice" package [18].

### Data enhancement

When data is imbalanced, the model performs better on types with higher distribution than on those with lower distribution. Therefore, it is necessary to deal with the unbalanced data. The ratio of the nDN group to the DN group in NPHDC data is 1.35: 1, which belongs to unbalanced data. We used the Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC) algorithm, designed for oversampling minority class samples, by randomly selecting a minority class sample as a starting point, choosing a neighboring sample as a reference point, and generating new synthetic samples between them to boost the minority class sample count ("themis" package) [19].

## Data splitting and normalization

The NPHDC data were randomly divided into the training set and test set by stratified sampling in the ratio of 7:3. The training set has a total of 2412 samples. The test set has a total of 1034 samples.

Since there were significant differences in the scales of the data in this dataset, making comparisons between different variables challenging, the min-max standardization method was applied to standardize the indicators in the training set. The maximum and minimum values of the training set were then used to standardize the test set and external validation data, ensuring uniform data standards.

## Balance test

Statistical testing methods were used to perform differential testing on the distribution of individual indicators in the training and test sets. For continuous variables, *t*-tests and Wilcoxon tests were applied based on whether the data met the assumptions of normality and homoscedasticity. The fold change (FC) was calculated to quantify differences between groups. For categorical variables, the chi-squared test or Fisher's exact test was conducted to compare inter-group differences, taking into account the total sample size and expected values. The odds ratio (OR) was used to assess the relative distribution of different groups.

To assess the overall balance of sample distribution, an analysis was conducted using Uniform Manifold Approximation and Projection (UMAP, "umap" package) and Permutation Multivariate Analysis of Variance (PERMANOVA, "vegan" package). The UMAP algorithm, a non-supervised dimensionality reduction technique based on manifold learning, maps high-dimensional data to a lower-dimensional space. Unlike linear methods such as Principal Component Analysis (PCA), UMAP preserves both the overall and local structures of the data [20]. PMANOVA is a non-parametric multivariate analysis of variance method based on F-statistics. It decomposes the total variance using distance matrices to test differences among multiple variables [21].

## Feature selection

Before constructing the model, feature selection for key variables is beneficial for reducing model complexity and enhancing predictive capability. In this study, we utilized four methods for feature selection: differential analysis, the Least Absolute Shrinkage and Selection Operator (LASSO, "glmnet" package), Recursive Feature Elimination (RFE, "caret" package), and the Max-Relevance and Min-Redundancy (MRMR, "mRMRe" package) algorithms. We used either *T*-tests or Wilcoxon tests for continuous variables and chi-squared tests or Fisher's exact tests for categorical variables, as previously described. FC or OR was used to characterize inter-group differences. LASSO regression is a regularization method based on linear regression. It utilizes L1 regularization as a penalty term, continually compressing the coefficients of variables, resulting in some coefficients becoming zero, thus achieving feature selection [22]. The RFE algorithm is a model-based feature selection method. It begins by training with all features, then recursively trains the model by adding or removing specific variables, eliminating the least important features to achieve feature selection [23]. MRMR is a feature selection method based on mutual information. It ensures maximal correlation of each variable with the occurrence of DN while minimizing redundancy among variables, effectively preventing significant collinearity between features [24].

## Construction and evaluation of machine learning and deep learning models

Using the key features selected as described, 8 machine learning models and 2 deep learning models were constructed.

Random Forest (RF) is an ensemble learning method. It makes predictions by constructing multiple decision trees, each trained on randomly selected subsets of data and features. The final output is determined by aggregating the predictions of all trees through voting ("randomForest" package) [25].

Logistic Regression (LR) is a linear model widely employed for classification problems. It assumes that the data follows a logistic distribution and utilizes maximum likelihood estimation to fit model parameters. Predictions are made by mapping the output of the linear function to probabilities within the [0, 1] range ("rms" package).

Naive Bayes (NB) is a class of probabilistic classifiers based on Bayes' theorem. It assumes that features are independent of each other and predicts classification by computing the conditional probability of features given each class ("e1071" package).

eXtreme Gradient Boosting (XGboost) is a gradient boosting algorithm. It iteratively trains multiple decision tree models, with each tree correcting the errors of the previous ones and optimizing the loss function to enhance model performance ("xgboost" package) [26].

Adaptive Boosting (Adaboost) is an iterative boosting algorithm. It sequentially trains a series of weak classifiers, adjusting the weights of previously misclassified samples in each iteration to improve model performance ("adabag" package) [27].

Category Boosting (Catboost) is a gradient boosting algorithm that can automatically handle categorical features.

It utilizes a histogram-based optimization method, eliminating the need for preprocessing such as one-hot encoding. Catboost can effectively handle large-scale datasets during training and boasts high accuracy and generalization capabilities ("catboost" package) [28].

Logit Boosting (Logitboost) is a boosting algorithm that employs optimizing logistic regression models to enhance performance. It minimizes the log loss function, progressively adds new models, and adjusts sample weights to improve model performance ("caTools" package).

Light Gradient Boosting Machine (LightGBM) is a tree-based gradient boosting framework that features efficiency and distributed training capabilities. It utilizes a histogram-based decision tree algorithm, enabling rapid tree model construction during training and efficient handling of large-scale datasets ("lightgbm" package) [29].

Feedforward Neural Network (FNN) is a classic neural network structure comprising input, hidden, and output layers. Information flows from the input layer through the hidden layers, ultimately outputting to the output layer. By training and adjusting weights, FNN learns the relationships between inputs and outputs ("keras", "tensorflow" packages) [30].

Back Propagation Neural Network (BPNN) is a classic neural network model that utilizes the backpropagation algorithm to adjust weights within the network, aiming to minimize the error between predicted outputs and actual outputs. It consists of input, hidden, and output layers, with information propagating from the input layer through the hidden layers to the output layer ("keras", "tensorflow" packages) [30].

All models were trained with default parameters and evaluated using 5-fold cross-validation to ensure the reliability of the results.. Subsequently, model reliability and accuracy were compared using Receiver Operating Characteristic (ROC) curves, Precision-Recall (PR) curves, Accuracy, Matthews Correlation Coefficient (MCC), and Kappa values. The best predictive model was selected based on these metrics, and it was tested on the test set.

Then, to ensure the selected RF model achieved optimal performance, we conducted hyperparameter tuning. We varied the number of features considered for each tree split (mtry) and the minimum size of each leaf node (nodesize) from 1 to 15. Additionally, we varied the number of trees from 100 to 1000 (ntrees, in increments of 25), resulting in a total of 8325 model configurations. The model with the highest ROC-AUC value was chosen as the final model for subsequent analysis.

### Model interpretability

Machine learning models are often considered black-box models because the mapping process from inputs to outputs is not readily observable, making it challenging to understand their internal workings. To enhance model interpretability and gain insight into how input features influence model outputs, SHAP values and Partial Dependence Plots (PDP) were employed for interpretability analysis.

### Statistical analysis

The study involved analysis using R (version 4.3.1) and Python (version 3.9.3). For continuous variables, group differences were compared using t test or Wilcoxon test. For categorical variables, group differences were assessed using chi-squared test or Fisher's exact test. Correlation analysis was conducted using the Spearman correlation test. Statistical differences were defined as $P < 0.05$.

## Results

### Training and test groups are well balanced

The difference analysis of each variable between the training group and the test group showed that only three variables had significant differences between the two groups, indicating reasonably good balance between the variables ($P < 0.05$, Fig. 1A). The UMAP and PERMANOVA analyses showed a more balanced distribution of the samples between the two groups ($P > 0.05$, Fig. 1B). Therefore, the training and testing groups were well balanced and could be used for subsequent studies.

### Feature selection

The differential analysis reveals that a total of 50 variables exhibit significant differences in distribution between the DN and nDN groups, including 31 continuous variables and 19 categorical variables ($P < 0.05$, Fig. 2A, Tables S1 and S2). The RFE results indicate that the RF model has the lowest error among the five models. Therefore, the RF-RFE method was selected, resulting in the selection of 45 key variables (Fig. 2B). In LASSO regression analysis, 47 key variables were selected when the λ value was minimized (Fig. 2C, D). The MRMR algorithm identified a total of 66 relevant variables. As shown in Fig. 2E, these four methods collectively identified 15 clinical indicators associated with DN, including Age, Body Mass Index (BMI), Diastolic Pressure (BP_LOW), Systolic Pressure (BP_HIGH), Total Cholesterol (TC), Blood Urea (BU), Lactate Dehydrogenase (LDHL), Aspartate Transaminase (AST), Gamma-Glutamyltransferase (GGT), Lipase (LPS), Albumin (ALB), Serum Creatinine (SCR), Fibrinogen (FBG), Direct Bilirubin (DBILI), and Albumin Creatinine Ratio (ALB_CR). Furthermore, the Spearman correlation

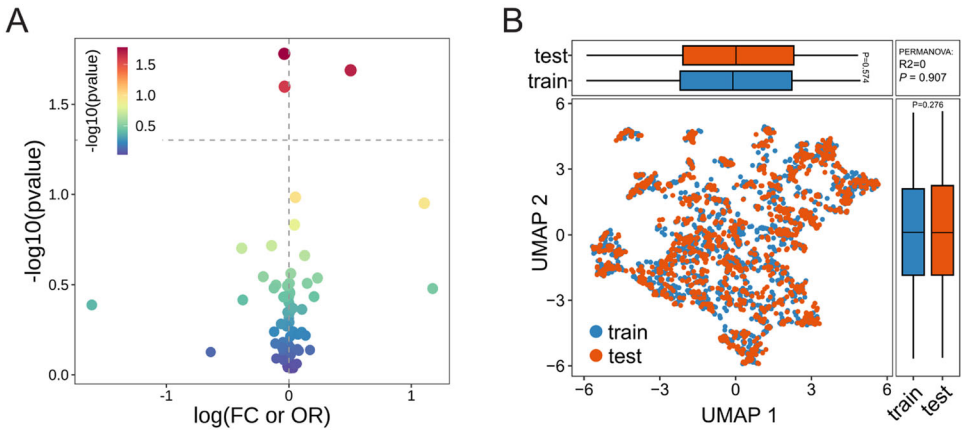


**Fig. 1** The balance test between the training and testing groups. **A** Volcano plot for single-variable difference analyses (The horizontal axis represents the FC or OR values of each variable between the training set and the test set, while the vertical axis represents the *P*-values.). **B** The scatter plot of UMAP and PERMANOVA analyses between the training set and the test set

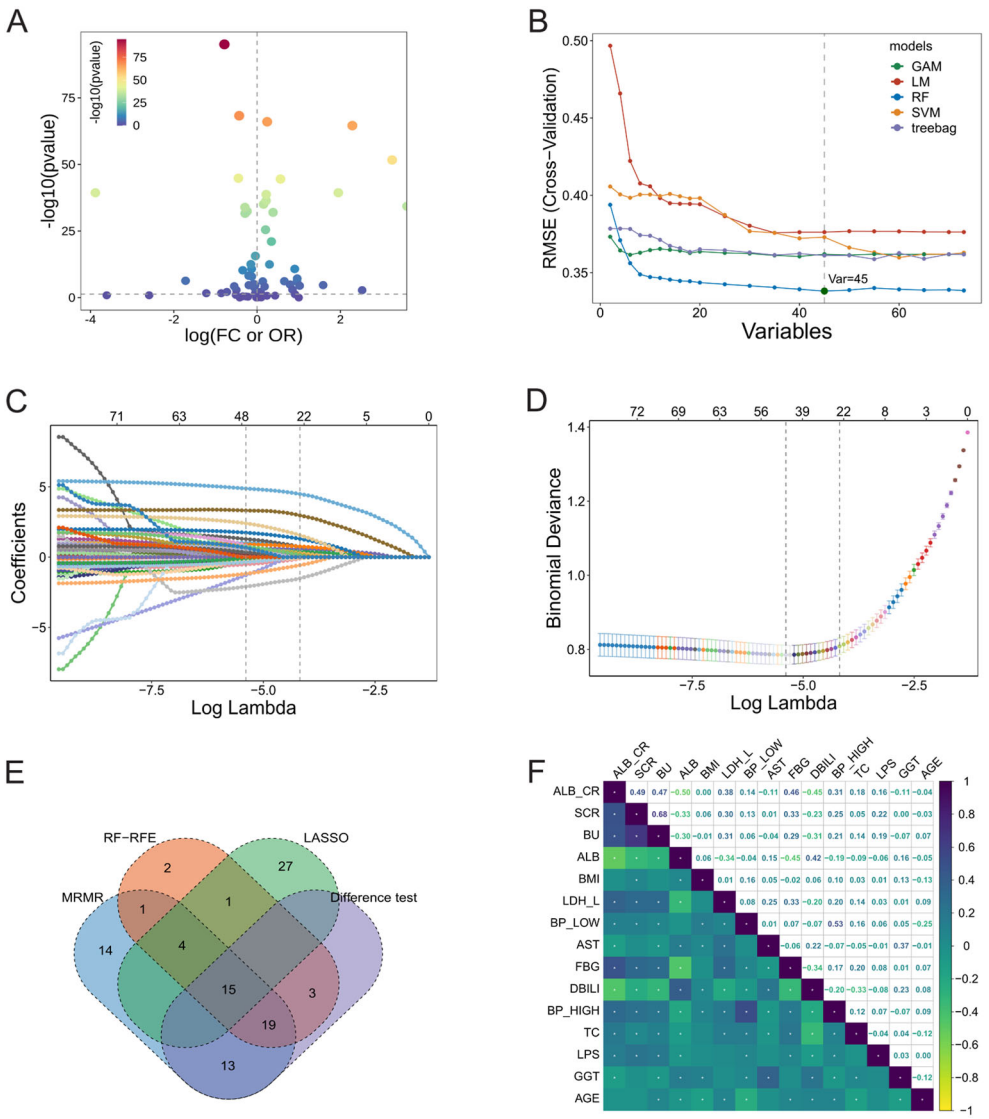


**Fig. 2** Screening of key DN features. **A** volcano plot for difference analysis. **B** RFE. Each variable's LASSO regression coefficient (**C**) and the binomial deviation results of Lasso regression through 10-fold cross-validation (**D**). **E** Venn diagram of the selection results from the four methods. **F** Heatmap of the correlations among key features (*$P < 0.05$)

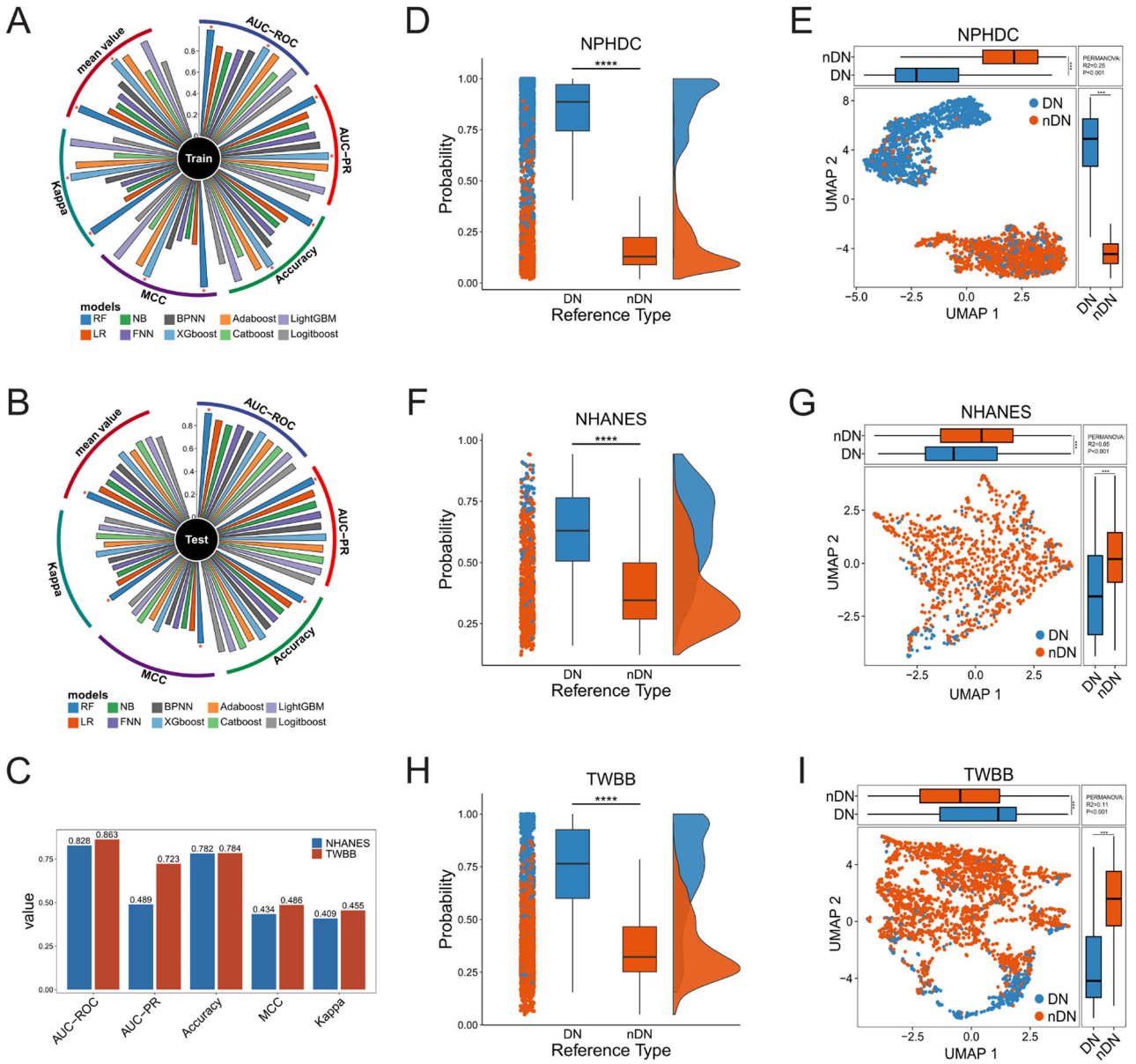


**Fig. 3** Construction and evaluation of the DN prediction model (***$P < 0.001$). Bar plots of each evaluation metric for the 10 models in the training set (**A**) and the test set (**B**) (* is the best model). **C** Validation of RF models in the NHANES and TWBB cohorts. Boxplots of the RF predicted probability in DN and nDN groups in the NPHDC (**D**), NHANES (**F**), and TWBB (**H**) cohorts, and scatter plots of the UMAP distributions based on the RF probability and the key features in the NPHDC (**E**), NHANES (**G**) and TWBB (**I**) cohorts

coefficients between these 15 variables are relatively small, indicating the absence of significant collinearity (Fig. 2F).

## Construction and evaluation of the DN prediction model

Among the 10 models, the RF and XGboost models displayed excellent predictive performance in all 5 evaluation metrics in the training set (all equal to 1, Fig. 3A, Table S3). In the test set, the RF model exhibited superior predictive capabilities compared to other models (AUC-ROC = 0.912, AUC-PR = 0.930, Accuracy = 0.852, MCC = 0.707, Kappa = 0.704, Fig. 3B, Table S3). Therefore, the RF model was selected as the final prediction model. Then, we performed hyperparameter tuning for the RF model. When mtry = 2, nodesize = 2, and ntrees = 425, the model achieved the best ROC-AUC value of 0.915. Compared to before tuning, the model's performance improved by 0.329% (Table S4). External validation results from the NHANES and TWBB cohorts indicated that the RF model demonstrated good generalization ability, with ROC-AUC of 0.828 and 0.863, reflecting strong predictive

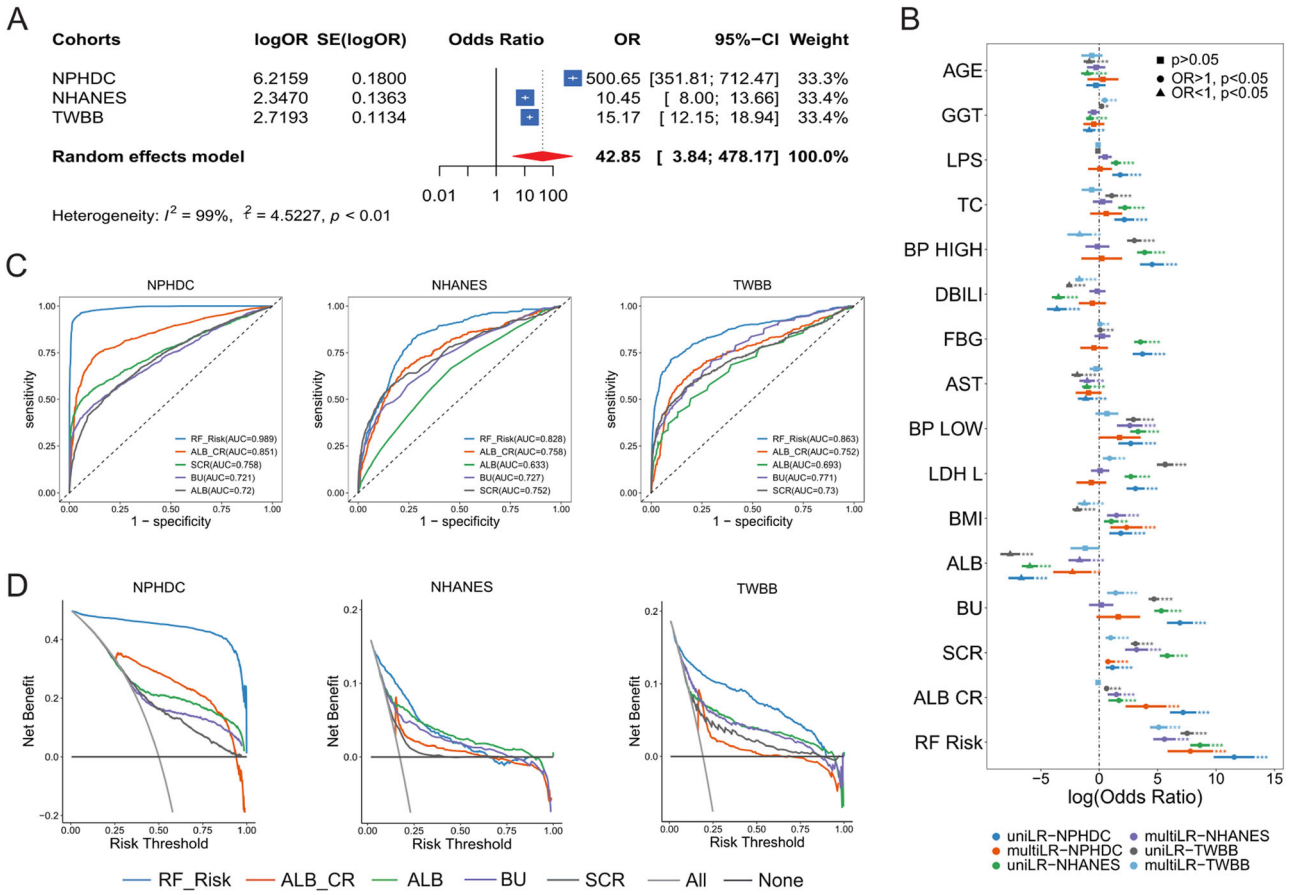


**Fig. 4** Comparison of the predictive value of the RF model with other indicators (***$P < 0.001 <$ **$P < 0.01 <$ *$P < 0.05$). **A** Meta-analysis; **B** logistic regression; **C** ROC curve; **D** DCA

performance. However, its performance in terms of AUC-PR, MCC, and Kappa is not satisfactory, which may be due to the imbalance in the validation data. (Fig. 3C). Furthermore, boxplots and UMAP results illustrated that the risk prediction outcomes based on this model could effectively differentiate between DN and nDN patients in the NPHDC, NHANES, and TWBB cohorts ($P < 0.05$, Fig. 3D–I).

## The RF model outperforms other diagnostic indicators

Meta-analysis showed that the RF model predicted risk as an adverse effect factor in the NPHDC cohort, NHANES cohort, and TWBB cohort (OR > 1, $P < 0.001$, Fig. 4A). In addition, we further compared the predictive ability of the RF model with other clinically used metrics for diagnosing DN by LR, ROC curve, and Decision Curve Analysis (DCA). In the 3 cohorts, the univariate and multivariate LR results showed that the OR of RF in predicting risk was greater than 1 and significantly higher than other indicators (Fig. 4B). The area under the curve of ROC was 0.989 in the NPHDC cohort, 0.828 in the NHANES data, and 0.863 in the TWBB cohort, which were higher than other clinical indices such as ALB_CR, SCR and BU (Fig. 4C). DCA results revealed that patients benefited significantly more from clinical treatment based on the RF risk prediction results compared to other indicators (Fig. 4D). In conclusion, the predictive power of RF is significantly better than that of the currently used diagnostic indicators of DN.

## Explanation analysis of the RF model

The importance of ALB_CR in the RF model is significantly higher than other indicators, followed by SCR and BU (Fig. 5A). Additionally, when the values of ALB_CR, SCR, and BU increase, the RF prediction probability also increases ($R > 0$). However, the value of ALB is significantly negatively correlated with the RF prediction probability ($R < 0$, Fig. 5B, C).

Furthermore, we found that the samples incorrectly predicted by the RF model in the 3 cohorts were mainly concentrated in the DN group (Fig. 5D–F), and their prediction probabilities were mainly distributed between 0.2 and 0.8 (Fig. 5G). So, when the model's prediction results fall within the intermediate range, clinical judgment is needed to assess the reliability of these results.

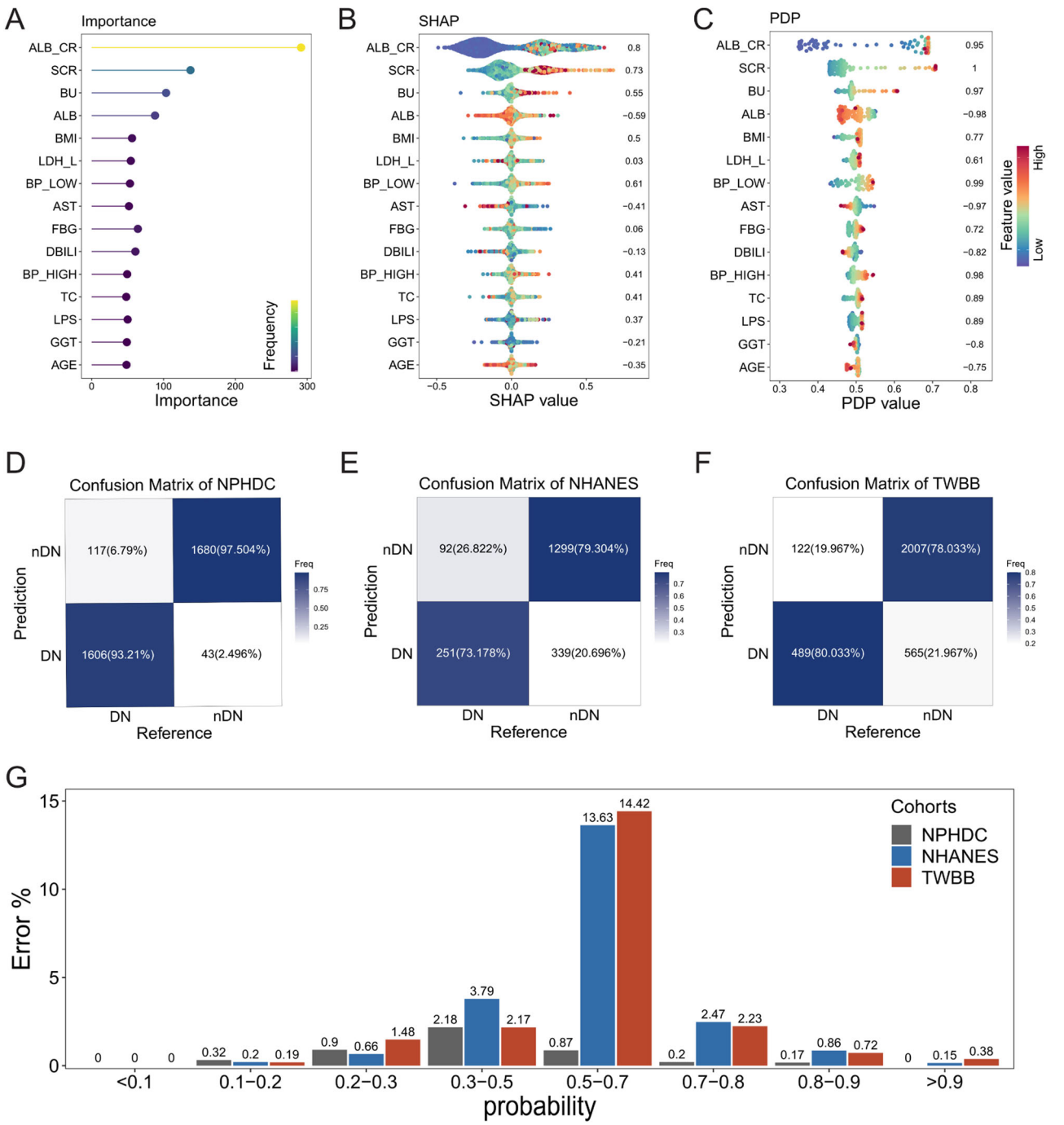


**Fig. 5** Explanation analysis of the RF model. **A** The importance of various indicators in the RF model. RF model interpretability based on SHAP values (**B**), and PDP (**C**). Confusion matrix of the NPHDC (**D**), NHANES (**E**), and TWBB (**F**) cohorts. **G** Percentage probability distribution of samples with RF prediction errors in the 3 cohorts

## Discussion

Due to the complex and insidious nature of DN pathogenesis, early diagnosis and treatment are challenging, and single clinical characteristic indicators typically struggle to make accurate diagnoses. Currently, clinical indicators such as proteinuria, creatinine, and glomerular filtration rate have some suggestive value in assessing the onset and prognosis of DN. However, due to the relatively complex mechanisms underlying DN pathogenesis and the relative independence of various risk indicators, it is difficult to make precise determinations about the occurrence and screening of DN [31].

With the advancement of computer science and the availability of large medical datasets, machine learning has

become closely integrated with the field of medicine. It has found widespread applications in various areas of research, including epidemiology, oncology, and immunological diseases [32, 33]. At present, the trend in clinical early diagnosis is the use of multiple indicators in combination. Constructing a DN risk prediction model by combining various clinical indicators is beneficial for disease prevention, diagnosis, and treatment. Therefore, in this study, we analyzed clinical data of diabetic patients using various machine learning algorithms to identify clinical indicators related to the occurrence of DN. Then, We built a DN prediction model based on these indicators and validated the model internally and externally. The goal is to provide assistance in the early diagnosis and prevention of DN for patients.

Our study identified 15 clinical indicators such as AGE, BMI, ALB_CR, SCR, BU, and ALB as factors associated with the occurrence of DN. Some of these indicators are recognized as common risk factors for type 2 diabetes, such as age and BMI, and these factors were also associated with the development of DN in our study [34]. Currently, ALB_CR and glomerular filtration rate are commonly used clinical indicators for early DN screening and have been included in expert consensus [35]. Another study found that increased levels of SCR led to a 30-fold increase in the risk of DN and a 6.5-fold increase in the risk of death, and SCR has now been used clinically as a diagnostic indicator for the progression of DN to end-stage renal disease [36]. Moreover, BU, ALB, and LDHL have all been shown to be strongly associated with the development of DN [37–39]. These findings further confirm the reliability and scientific validity of our study.

We have successfully constructed a risk prediction model for DN based on these indicators. In our study, we found that the traditional Random Forest model achieved the best predictive performance, surpassing not only other machine learning algorithms but also two deep learning models. As we know, deep learning models have more complex algorithms and advanced learning capabilities compared to traditional machine learning models. However, they do depend on extensive datasets to complete the learning process [40]. One of the possible reasons for the poor performance of deep learning models in this study could be the relatively small training sample size, which only consists of 2412 cases.

In the RF model, the levels of ALB_CR, SCR, and BU are positively correlated with the model's predicted probability, whereas the level of ALB is negatively correlated. An increase in the levels of ALB_CR, SCR, and BU is observed during the development of DN, but the ALB levels may not always increase. Based on the ALB levels in urine, DN can be divided into proteinuria phenotype and non-proteinuria phenotype. In patients with these two different phenotypes, there are morphological and functional differences [41]. This could explain why the ALB levels are negatively correlated with the RF prediction values.

In conclusion, our study successfully developed a risk prediction model for diagnosing DN, which demonstrated good performance in both the internal training and test sets, as well as in external validation cohorts. This finding provides a powerful tool for early diagnosis and intervention of DN, with the potential to reduce disease progression and its adverse impact on patients. However, the study has some limitations. Firstly, although our model showed good performance, there is still a certain error rate, especially when predicting probabilities close to intermediate values. Secondly, this is a case-control study, and it may have some selection bias, which may not accurately reflect the real population with DN. Additionally, the study only analyzed data from Chinese and American populations and had limited sources for validation. In the future, we plan to incorporate more biological markers and genomic information to further improve the accuracy of risk prediction.

## Conclusion

We successfully identified 15 clinical features that are closely related to the occurrence of DN and constructed an RF model based on these features, which has good predictive ability in the training set as well as in 2 external validation cohorts. The model is expected to be useful for early screening of clinical DN patients.

### Data availability

All data in this article can be found in the following databases: NPHDC, NHANES, and TWBB. An online platform has been created and you can access it through the following link (https://dn-prediction.shinyapps.io/DN-PRED-English).

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1007/s12020-024-03735-1.

**Acknowledgements** We thank the National Population Health Data Center of China, the National Health and Nutrition Examination Survey of the United States, and Taiwan Biobank for providing data support.

**Author contributions** All authors contributed to the study's conception and design. Data collection and analysis were performed by J.J.M. The first draft of the manuscript was written by J.J.M., S.G.A., and M.H.C. The revision of the manuscript was completed by L.Z. and J.L. All authors read and approved the final manuscript.

**Funding** This study was supported by the College Students' Innovative Entrepreneurial Training Plan Program (202310367071).

### Compliance with ethical standards

**Conflict of interest** The authors declare no competing interests.

## References


1. M. Darenskaya, S. Kolesnikov, N. Semenova, L. Kolesnikova. Diabetic nephropathy: significance of determining oxidative stress and opportunities for antioxidant therapies. Int. J. Mol. Sci. 24 (2023). https://doi.org/10.3390/ijms241512378.
2. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017, A systematic analysis for the Global Burden of Disease Study 2017. Lancet **392**, 1789–1858 (2018). https://doi.org/10.1016/s0140-6736(18)32279-7
3. M. Guedes, R. Pecoits-Filho, Can we cure diabetic kidney disease? Present and future perspectives from a nephrologist's point of view. J. Intern. Med. **291**, 165–180 (2022). https://doi.org/10.1111/joim.13424
4. Q. Hu, Y. Chen, X. Deng, Y. Li, X. Ma, J. Zeng, Y. Zhao, Diabetic nephropathy: Focusing on pathological signals, clinical treatment, and dietary regulation. Biomed. Pharmacother. **159**, 114252 (2023). https://doi.org/10.1016/j.biopha.2023.114252
5. K. Zhang, Z. Fu, Y. Zhang, X. Chen, G. Cai, Q. Hong, The role of cellular crosstalk in the progression of diabetic nephropathy. Front. Endocrinol. (Lausanne) **14**, 1173933 (2023). https://doi.org/10.3389/fendo.2023.1173933
6. M. Vučić Lovrenčić, S. Božičević, L. Smirčić Duvnjak, Diagnostic challenges of diabetic kidney disease. Biochem. Med. (Zagreb) **33**, 030501 (2023). https://doi.org/10.11613/bm.2023.030501
7. R.Y. Choi, A.S. Coyner, J. Kalpathy-Cramer, M.F. Chiang, J.P. Campbell, Introduction to machine learning, neural networks, and deep learning. Transl. Vis. Sci. Technol. **9**, 14 (2020). https://doi.org/10.1167/tvst.9.2.14
8. G.S. Handelman, H.K. Kok, R.V. Chandra, A.H. Razavi, M.J. Lee, H. Asadi, eDoctor: Machine learning and the future of medicine. J. Intern Med. **284**, 603–619 (2018). https://doi.org/10.1111/joim.12822
9. R. Gupta, S. Kumari, A. Senapati, R.K. Ambasta, P. Kumar, New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. Ageing Res. Rev. **90**, 102013 (2023). https://doi.org/10.1016/j.arr.2023.102013
10. Z. Bao, J. Bufton, R.J. Hickman, A. Aspuru-Guzik, P. Bannigan, C. Allen, Revolutionizing drug formulation development: The increasing impact of machine learning. Adv. Drug Deliv. Rev. **202**, 115108 (2023). https://doi.org/10.1016/j.addr.2023.115108
11. J.B. Xue, S. Xia, X.Y. Wang, L.L. Huang, L.Y. Huang, Y.W. Hao, L.J. Zhang, S.Z. Li, Recognizing and monitoring infectious sources of schistosomiasis by developing deep learning models with high-resolution remote sensing images. Infect. Dis. Poverty **12**, 6 (2023). https://doi.org/10.1186/s40249-023-01060-9
12. J.M. Yin, Y. Li, J.T. Xue, G.W. Zong, Z.Z. Fang, L. Zou, Explainable machine learning-based prediction model for diabetic nephropathy. J. Diabetes Res. **2024**, 8857453 (2024). https://doi.org/10.1155/2024/8857453
13. M. Xu, H. Zhou, P. Hu, Y. Pan, S. Wang, L. Liu, X. Liu, Identification and validation of immune and oxidative stress-related diagnostic markers for diabetic nephropathy by WGCNA and machine learning. Front. Immunol. **14**, 1084531 (2023). https://doi.org/10.3389/fimmu.2023.1084531
14. X.Z. Liu, M. Duan, H.D. Huang, Y. Zhang, T.Y. Xiang, W.C. Niu, B. Zhou, H.L. Wang, T.T. Zhang, Predicting diabetic kidney disease for type 2 diabetes mellitus by machine learning in the real world: A multicenter retrospective study. Front Endocrinol. (Lausanne) **14**, 1184190 (2023). https://doi.org/10.3389/fendo.2023.1184190
15. S.M. Hosseini Sarkhosh, M. Hemmatabadi, A. Esteghamati, Development and validation of a risk score for diabetic kidney disease prediction in type 2 diabetes patients: a machine learning approach. J. Endocrinol. Invest **46**, 415–423 (2023). https://doi.org/10.1007/s40618-022-01919-y
16. L. Zhao, H. Ren, J. Zhang, Y. Cao, Y. Wang, D. Meng, Y. Wu, R. Zhang, Y. Zou, H. Xu et al. Diabetic retinopathy, classified using the lesion-aware deep learning system, predicts diabetic end-stage renal disease in Chinese patients. Endocr. Pract. **26**, 429–443 (2020). https://doi.org/10.4158/ep-2019-0512
17. C.T. Fan, J.C. Lin, C.H. Lee, Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. Pharmacogenomics **9**, 235–246 (2008). https://doi.org/10.2217/14622416.9.2.235
18. S.v. Buuren. *Flexible Imputation of Missing Data*, 2nd edn. (Boca Raton, FL, 2018)
19. Z. Xu, D. Shen, Y. Kou, T. Nie. A synthetic minority oversampling technique based on Gaussian mixture model filtering for imbalanced data classification. *IEEE Trans Neural Netw Learn Syst* (2022). https://doi.org/10.1109/tnnls.2022.3197156
20. L. McInnes, J. Healy, J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 2018,
21. J. Chen, X. Zhang, D-MANOVA: fast distance-based multivariate analysis of variance for large-scale microbiome association studies. Bioinformatics **38**, 286–288 (2021). https://doi.org/10.1093/bioinformatics/btab498
22. J.K. Tay, B. Narasimhan, T. Hastie. Elastic net regularization paths for all generalized linear models. J. Stat. Softw. 106 (2023). https://doi.org/10.18637/jss.v106.i01
23. Y. Han, L. Huang, F. Zhou, A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. Bioinformatics **37**, 2183–2189 (2021). https://doi.org/10.1093/bioinformatics/btab055
24. H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005). https://doi.org/10.1109/tpami.2005.159
25. L. Breiman, Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
26. T. Chen, C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794 (2016)
27. E. Alfaro, M. Gáamez, N. García. adabag: An R package for classification with boosting and bagging. J. Stat. Softw. 2013, 54, https://doi.org/10.18637/jss.v054.i02
28. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin. CatBoost: unbiased boosting with categorical features. Adv. Neural Inform. Process. Syst. **31** (2018). https://doi.org/10.48550/arXiv.1706.09516
29. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA*, 2017; pp. 3149–3157
30. M. Abadi, P. Barham, J. Chen, Z. Chen, X. Zhang. TensorFlow: A system for large-scale machine learning. *USENIX Association* 2016, 265–283, https://doi.org/10.48550/arXiv.1605.08695
31. T.A. Dejenie, E.C. Abebe, M.A. Mengstie, M.A. Seid, N.A. Gebeyehu, G.A. Adella, G.A. Kassie, A.Y. Gebrekidan, M.M. Gesese, K.D. Tegegne et al. Dyslipidemia and serum cystatin C levels as biomarker of diabetic nephropathy in patients with type 2 diabetes mellitus. Front Endocrinol. (Lausanne) **14**, 1124367 (2023). https://doi.org/10.3389/fendo.2023.1124367
32. A.K. Clift, D. Dodwell, S. Lord, S. Petrou, M. Brady, G.S. Collins, J. Hippisley-Cox, Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. Bmj **381**, e073800 (2023). https://doi.org/10.1136/bmj-2022-073800

33. V. Subbiah, The next generation of evidence-based medicine. Nat. Med **29**, 49–58 (2023). https://doi.org/10.1038/s41591-022-02160-z
34. R.D. Joshi, C.K. Dhakal. Predicting type 2 diabetes using logistic regression and machine learning approaches. Int. J. Environ. Res. Public Health **18** (2021). https://doi.org/10.3390/ijerph18147346
35. A. Zanchi, A.W. Jehle, F. Lamine, B. Vogt, C. Czerlau, S. Bilz, H. Seeger, S. de Seigneux, Diabetic kidney disease in type 2 diabetes: a consensus statement from the Swiss Societies of Diabetes and Nephrology. Swiss Med Wkly **153**, 40004 (2023). https://doi.org/10.57187/smw.2023.40004
36. B.F. Palmer, Change in albuminuria as a surrogate endpoint for cardiovascular and renal outcomes in patients with diabetes. Diabetes Obes. Metab. **25**, 1434–1443 (2023). https://doi.org/10.1111/dom.15030
37. X. Ren, N. Kang, X. Yu, X. Li, Y. Tang, J. Wu, Prevalence and association of diabetic nephropathy in newly diagnosed Chinese patients with diabetes in the Hebei province: A single-center case-control study. Medicine (Baltimore) **102**, e32911 (2023). https://doi.org/10.1097/md.0000000000032911
38. S. Chen, L. Chen, H. Jiang, Prognosis and risk factors of chronic kidney disease progression in patients with diabetic kidney disease and non-diabetic kidney disease: a prospective cohort CKD-ROUTE study. Ren. Fail **44**, 1309–1318 (2022). https://doi.org/10.1080/0886022x.2022.2106872
39. K. Azushima, J.P. Kovalik, T. Yamaji, J. Ching, T.W. Chng, J. Guo, J.J. Liu, M. Nguyen, R.B. Sakban, S.E. George, et al. Abnormal lactate metabolism is linked to albuminuria and kidney injury in diabetic nephropathy. Kidney Int. (2023). https://doi.org/10.1016/j.kint.2023.08.006
40. J.G. Greener, S.M. Kandathil, L. Moffat, D.T. Jones, A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. **23**, 40–55 (2022). https://doi.org/10.1038/s41580-021-00407-0
41. M. Garofolo, V. Napoli, D. Lucchesi, S. Accogli, M.L. Mazzeo, P. Rossi, E. Neri, S. Del Prato, G. Penno, Type 2 diabetes albuminuric and non-albuminuric phenotypes have different morphological and functional ultrasound features of diabetic kidney disease. Diabetes Metab. Res Rev. **39**, e3585 (2023). https://doi.org/10.1002/dmrr.3585