BMJ Open

Development and internal validation of an interpretable risk prediction model for diabetic peripheral neuropathy in type 2 diabetes: a single-centre retrospective cohort study in China

Lianhua Liu ,[1] Bo Bi,[1] Mei Gui,[1] Linli Zhang,[2] Feng Ju,[3] Xiaodan Wang,[1] Li Cao[1]

To cite: Liu L, Bi B, Gui M, *et al*. Development and internal validation of an interpretable risk prediction model for diabetic peripheral neuropathy in type 2 diabetes: a single-centre retrospective cohort study in China. *BMJ Open* 2025;**15**:e092463. doi:10.1136/bmjopen-2024-092463

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-092463).

Received 15 August 2024
Accepted 07 March 2025

[1]Department of Biostatistics, School of Public Health, Hainan Medical University, Haikou, Hainan, China
[2]Department of Mathematics, Physics, and Chemistry teaching, Hainan University, Haikou, Hainan, China
[3]Department of Endocrinology, The Second Affiliated Hospital of Hainan Medical University, Haikou, Hainan, China

Correspondence to
Professor Li Cao;
caoli@hainmc.edu.cn

ABSTRACT

Objective Diabetic peripheral neuropathy (DPN) is a common and serious complication of diabetes, which can lead to foot deformity, ulceration, and even amputation. Early identification is crucial, as more than half of DPN patients are asymptomatic in the early stage. This study aimed to develop and validate multiple risk prediction models for DPN in patients with type 2 diabetes mellitus (T2DM) and to apply the Shapley Additive Explanation (SHAP) method to interpret the best-performing model and identify key risk factors for DPN.

Design A single-centre retrospective cohort study.

Setting The study was conducted at a tertiary teaching hospital in Hainan.

Participants and methods Data were retrospectively collected from the electronic medical records of patients with diabetes admitted between 1 January 2021 and 28 March 2023. After data preprocessing, 73 variables were retained for baseline analysis. Feature selection was performed using univariate analysis combined with recursive feature elimination (RFE). The dataset was split into training and test sets in an 8:2 ratio, with the training set balanced via the Synthetic Minority Over-sampling Technique. Six machine learning algorithms were applied to develop prediction models for DPN. Hyperparameters were optimised using grid search with 10-fold cross-validation. Model performance was assessed using various metrics on the test set, and the SHAP method was used to interpret the best-performing model.

Results The study included 3343 T2DM inpatients, with a median age of 60 years (IQR 53–69), and 88.6% (2962/3343) had DPN. The RFE method identified 12 key factors for model construction. Among the six models, XGBoost showed the best predictive performance, achieving an area under the curve of 0.960, accuracy of 0.927, precision of 0.969, recall of 0.948, F1-score of 0.958 and a G-mean of 0.850 on the test set. The SHAP analysis highlighted C reactive protein, total bile acids, gamma-glutamyl transpeptidase, age and lipoprotein(a) as the top five predictors of DPN.

Conclusions The machine learning approach successfully established a DPN risk prediction model with excellent performance. The use of the interpretable SHAP method could enhance the model's clinical applicability.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study is based on a large, high-dimensional dataset, enhancing the robustness of the findings.
⇒ By using multiple machine learning algorithms combined with techniques such as recursive feature elimination and Synthetic Minority Over-sampling Technique, comprehensive model development was ensured, resulting in strong predictive performance.
⇒ The Shapley Additive Explanation method enhanced model interpretability, identified key risk factors for diabetic peripheral neuropathy and improved clinical applicability.
⇒ The single-centre setting may limit the generalisability of the study results.

INTRODUCTION

According to the International Diabetes Federation's Diabetes Atlas (10th edition), 536.6 million adults worldwide had diabetes in 2021. About 6.7 million deaths were attributed to diabetes and its complications. China is projected to have 174.4 million people with diabetes by 2045, ranking first globally.[1] Diabetic peripheral neuropathy (DPN) is one of the most common and debilitating chronic complications of diabetes,[2 3] affecting over 50% of patients with type 2 diabetes mellitus (T2DM).[4–6] DPN, which results from prolonged hyperglycaemia, can lead to severe consequences such as foot deformities, ulceration and even amputation.[6–8] Each week in England, there are about 169 amputations in people with diabetes, and almost all of these individuals have DPN.[3] DPN can also damage the central nervous system and increase the risk of all-cause and cardiovascular mortality.[9] Despite its severe consequences, over 50% of patients with DPN remain asymptomatic in the early stage,[3] often leading to delayed diagnosis. Once symptoms like numbness and pain in

the limbs emerge, the nerve damage is typically irreversible,[2] posing a significant burden on both patients and the healthcare system.[3 10] Therefore, early screening and diagnosis are crucial. Although nerve conduction examinations are currently the gold standard for diagnosing DPN, they are labour-intensive, time-consuming and costly.[3] Given the large patient population, these examinations are impractical for routine screening in clinical practice, particularly in primary care settings where there is limited access to specialised equipment. Therefore, it is necessary to establish a DPN risk prediction model to identify high-risk patients for targeted screening and early diagnosis. Furthermore, since the factors contributing to the pathogenesis of DPN are not fully understood,[11] factor analysis can help clinicians in implementing targeted, personalised interventions and treatments. Such approaches may prevent or delay the onset of DPN and its complications, ultimately improving patients' quality of life and reducing the burden of disease on society.

Although extensive studies have explored DPN prediction, most rely on traditional single-algorithm approaches such as nomogram[12] and logistic regression (LR).[11 13] However, these conventional and singular methods may only capture limited data patterns, potentially reducing the accuracy and reliability of predictions. With the rapid advancements in artificial intelligence, machine learning (ML) is playing an increasingly important role in biomedical research, personalised medicine and computer-aided diagnosis.[14 15] Its applications in diabetes-related prediction research are also remarkably extensive.[9 16–20] Several researchers have used ML to develop DPN prediction models.[2 9 21–25] Previous studies have laid the foundation for this research; however, these studies either used a single method for modelling, or only focused on painful DPN, or were affected by small sample size, which in turn compromises performance. Most importantly, many studies suffered from the 'black-box' nature of the machine-learning models. Since it is difficult to explain which specific characteristics of patients lead to a particular prediction, this limits their clinical applicability.[26 27] To address these challenges, this study proposes developing an optimised DPN prediction model using a large-scale dataset. Furthermore, we propose to interpret the model through Shapley Additive Explanation (SHAP),[28] a game-theory-derived framework, to quantify feature contributions and elucidate the prediction mechanism. This approach will enhance the transparency and clinical utility of the model.

The aim of this study was to develop and validate a DPN prediction model for patients with T2DM, enabling clinicians to easily and accurately identify high-risk patients for further diagnosis and treatment. To enhance the model's predictive performance, we employed the recursive feature elimination (RFE) method[29] for optimal feature selection, the Synthetic Minority Over-sampling Technique (SMOTE)[30] for data balancing, and six widely used ML algorithms to construct the models, ultimately selecting the best-performing one. Furthermore, we employed the SHAP method to interpret the best-performing model, thereby improving its clinical applicability.

MATERIALS AND METHODS

Data collection

We retrieved data from 4372 hospitalised patients with diabetes admitted to the Department of Endocrinology at a tertiary hospital in Hainan between 1 January 2021 and 28 March 2023. The inclusion criteria were: (1) patients diagnosed with T2DM and (2) age ≥18 years. The exclusion criteria included: (1) neuropathy caused by other reasons, such as cervical or lumbar spondylopathy; (2) acute complications of diabetes, such as diabetic ketoacidosis and (3) other serious illnesses, such as malignant tumours or heart failure. To enhance the performance of ML models in this study, we included all patients who met the inclusion and exclusion criteria to ensure comprehensive model training. Finally, a total of 3346 patients were included in the final analysis. Patients were categorised into DPN and non-DPN groups based on discharge diagnosis records made from the results of nerve conduction examinations. Figure 1 outlines the overall research process.

Data extraction

The raw data retrieved from the hospital information system included more than 100 variables, encompassing demographic characteristics of the patients, laboratory results during hospitalisation, and comorbidity and complication information from discharge diagnosis records. Based on advice from clinicians and a review of relevant literature,[2 7–9 12 22 23 31–34] 80 potential variables (see online supplemental materials 1) that may be related to DPN were selected for analysis. Demographic variables included age, height, weight and blood pressure (both systolic and diastolic). Comorbidity and complications considered included hypertension, diabetic retinopathy (DR), diabetic nephropathy, peripheral vascular disease (PVD) and others. The laboratory data included haemoglobin, glycated haemoglobin, total bile acids (TBA), total cholesterol, triglycerides, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, serum uric acid, total bilirubin, direct bilirubin, alkaline phosphatase, C reactive protein (CRP), free triiodothyronine (FT3) and other essential laboratory characteristics.

Data preprocessing

Data preparation is crucial in ML model development. For the initial dataset containing 3346 patients and 80 indicators, we first excluded patients who had more than 30% missing values, ultimately retaining 3343 patients, including 2962 DPN patients and 381 non-DPN patients. Next, we eliminated indicators with a missing rate exceeding 25%, except for CRP, which was retained due to its relevance to DPN. This process resulted in 73 variables being retained for subsequent analysis. After

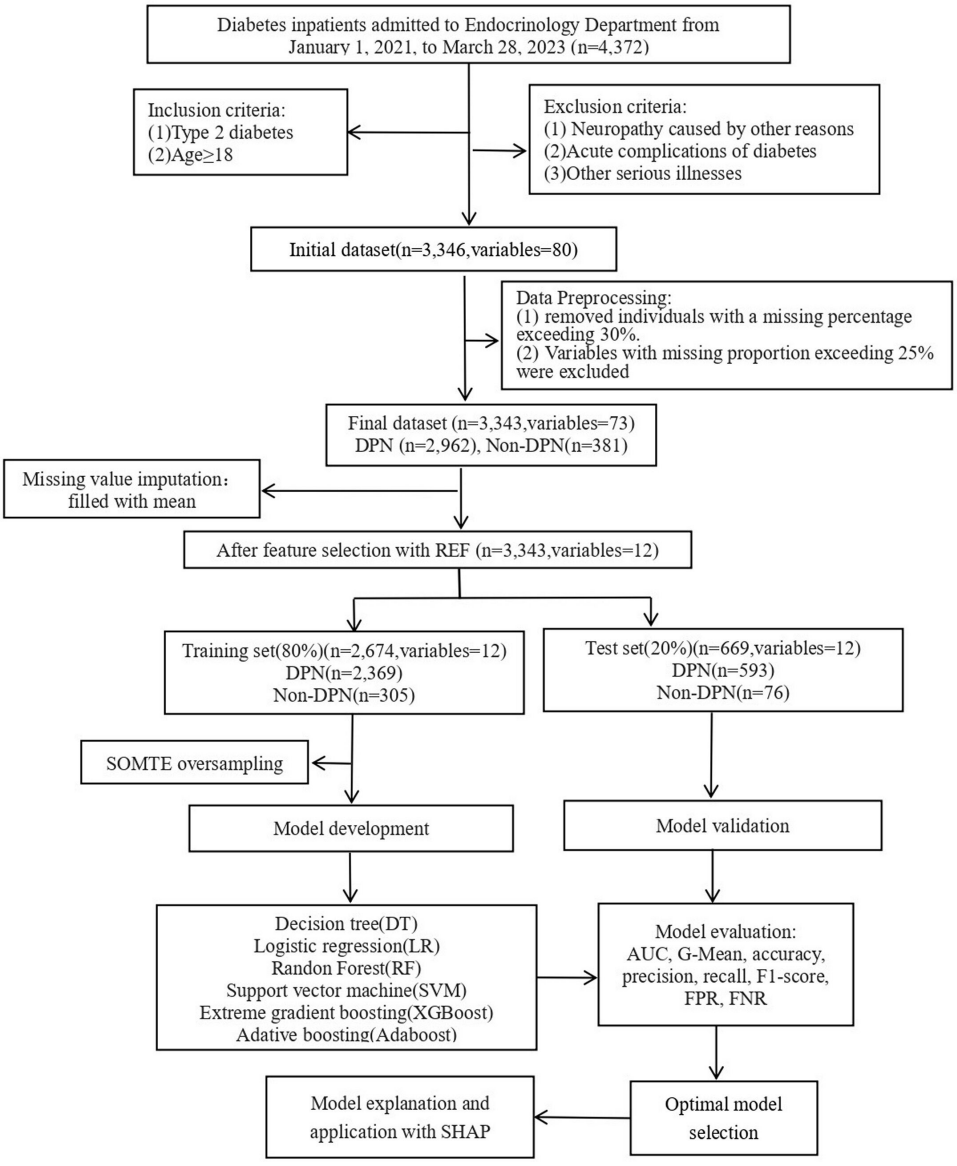

Figure 1 Study workflow. AUC, area under the curve; DPN, diabetic peripheral neuropathy; FNR, false negative rate; FPR, false positive rate; REF, recursive feature elimination; SHAP, Shapley Additive Explanation; SMOTE, Synthetic Minority Oversampling Technique.

eliminating variables, the remaining variables with missing values were imputed using the mean of the corresponding group.[9] The specific threshold used for excluding data was selected based on experiments and literature review.[9 19 22 35]

Feature selection

Continuous variables with skewed distributions were represented as M (P25, P75) and compared using the Mann-Whitney U test. Categorical variables were presented as n (%) and compared using the χ^2 test. Variables with a two-tailed $p<0.05$ were considered statistically significant and included in subsequent feature selection. To determine the optimal number of features for constructing the DPN prediction model, a random forest (RF)-based RFE approach was employed. This approach, when combined with classification algorithms, can identify a critical set of features and improve prediction performance. In this study, the RFE method was used in conjunction with six ML algorithms and fivefold cross-validation. After evaluating the performance of various classification algorithms based on the area under the curve (AUC) criterion to ascertain the optimal feature count, the most significant feature set can be identified.

Model construction

Based on the RFE[29] method, we selected the dataset corresponding to the best feature subset. The dataset was then divided into a training set (for model development) and a test set (for evaluation) in an 8:2 ratio, while preserving the balance of positive and negative samples. The dataset was derived from real-world hospitalised diabetes patients at a tertiary hospital, where most patients had more severe conditions, resulting in a higher number of DPN patients

compared with non-DPN patients. This created a highly imbalanced dataset. If left unadjusted, the minority class could be overwhelmed by the majority class, leading to suboptimal performance, as general ML algorithms often struggle with imbalanced datasets and tend to be biased towards the majority class.[36] To address this issue, we applied the SMOTE[30] method to generate synthetic samples for the minority class, thereby balancing the training set.

Subsequently, since different ML algorithms can learn different patterns of the data and each has its own advantages and disadvantages, we employed six well-known ML classifiers, which are widely used in the medical field, to develop prediction models on the balanced training set. These classifiers include decision trees,[37] LR,[38] RF,[39] support vector machines,[40] extreme gradient boosting (XGBoost)[41] and adaptive boosting (AdaBoost).[42] A grid search combined with 10-fold cross-validation was employed to optimise the hyperparameters of the six ML models to achieve the best AUC score.[43]

Model evaluation

To enhance the credibility of the model, we compared the performance of the six ML models using several evaluation metrics on the test set. These metrics included the AUC, G-mean, accuracy, precision, recall and F1-score. Additionally, we considered the false positive rate (FPR) and false negative rate (FNR), as these metrics play an important role in medical applications. A high FPR could result in unnecessary follow-up examinations, while a high FNR might lead to delayed treatment for patients, as their disease remains undiagnosed.

Model explanation

Prediction models constructed by ML algorithms are often complex and lack interpretability, commonly referred to as 'black-box' models, which can restrict their application in clinical settings. To address this issue, we used the SHAP approach for post hoc explainability analysis of the best-performing ML model. Grounded in game theory, SHAP values provide a measure of the impact each feature has on the model's predictions. This is achieved by assessing the change in predictions when a feature is included or excluded. These SHAP values help illustrate how much each predictor contributes to the model and whether the contribution is positive or negative, thereby enhancing the clarity and understanding of the model's decision-making process.[44]

Software and package application for modelling

All statistical analyses were conducted using IBM SPSS Statistics V.22.0. The establishment and interpretation of ML models were implemented in Python V.3.9 using scikit-learn V.1.2.2.

RESULTS

Baseline characteristics

A total of 3343 patients with T2DM were included in this study, with 2962 patients having DPN, showing a prevalence of 88.6%. The median age of the participants was 60.00 (IQR 53.00–69.00) years. The occurrence of PVD, DR, hypertension and cerebral infarction was significantly higher in the DPN group compared with the non-DPN group ($p<0.001$). Additionally, baseline measurements such as age, CRP, lipoprotein(a) (Lp-a), neutrophil-to-lymphocyte ratio (NLR) and 24-hour urine protein (24-hour UPr) were higher in DPN patients compared with non-DPN patients ($p<0.05$). Table 1 presents the baseline characteristics with significant differences.

Feature selection

We included 44 variables, which were identified as significant in univariate analysis, in the RFE method. The AUC trend with varying numbers of features for each classification algorithm is depicted in figure 2. According to the results, 12 variables were selected for further model development. They were age, weight, body mass index (BMI), glycated serum protein (GlySP), TBA, gamma-glutamyl transpeptidase (GGT), lactate dehydrogenase (LDH), Lp-a, NLR, 24 hours UPr, FT3 and CRP.

Modelling and evaluation

The original training set consisted of a total of 2369 cases of DPN and 305 cases of non-DPN. To address the class imbalance, the SMOTE method was applied, equalising the number of cases to 2369 for each category. The test set comprised a total of 593 cases of DPN and 76 cases of non-DPN. Six predictive models were developed based on the balanced training set, and their performance was evaluated using the test set. The employed hyperparameters for each ML model are presented in online supplemental table 1. The performance of the six established ML models on the test set is presented in table 2, figures 2 and 3. We can see that the XGBoost model exhibited the best performance, achieving the highest AUC (0.960), G-mean (0.850), accuracy (0.927), precision (0.969), recall (0.948) and F1-score (0.958). Additionally, the XGBoost model demonstrated the lowest FNR (0.052) and FPR (0.237). Therefore, the XGBoost model was selected for further predictive analysis and SHAP-based interpretation.

Model explanation

Online supplemental figure 1 illustrates the feature importance rankings determined by four different ML models. Despite some differences in their rankings, it is clear that variables such as age, GGT, CRP, TBA, 24-hour UPr and Lp-a play a crucial role.

We used the SHAP method to interpret the XGBoost model, which exhibited optimal performance on the test set. Figure 4 (left) illustrates the effect (positive or negative) of various factors on the XGBoost model. These features are listed in descending order of influence, with

Table 1 Univariate analysis results between the DPN group and non-DPN group (significant characteristics)

Characteristics	Total (n=3343)	Non-DPN (n=381)	DPN (n=2962)	P value
Age (years)	60.00 (53.00–69.00)	53.00 (43.00–62.00)	61.00 (54.00–70.00)	<0.001
Height (cm)	163.54 (158.00–170.00)	164.15 (160.00–170.00)	163.54 (158.00–170.00)	0.014
Weight (kg)	63.46 (56.00–70.00)	67.50 (57.00–74.00)	63.46 (55.50–69.53)	<0.001
BMI (kg/m^2)	23.73 (21.48–25.34)	24.68 (22.06–26.67)	23.73 (21.47–25.22)	<0.001
Diastolic blood pressure (mm Hg)	78.00 (70.00–85.00)	80.00 (73.00–87.50)	77.98 (70.00–85.00)	<0.001
Glycated serum protein	2.48 (2.11–2.71)	2.43 (2.05–2.73)	2.48 (2.12–2.71)	0.013
Total bile acids (µmol/L)	5.80 (3.00–7.78)	5.30 (2.65–5.90)	5.90 (3.10–7.78)	<0.001
Cystatin (mg/L)	0.86 (0.69–1.10)	0.80 (0.64–0.98)	0.87 (0.70–1.11)	<0.001
Gamma-glutamyl transferase (U/L)	33.00 (19.00–40.50)	40.00 (21.00–55.76)	32.00 (19.00–40.50)	<0.001
Alanine aminotransferase (U/L)	20.00 (14.00–30.00)	23.00 (15.00–40.00)	19.00 (14.00–29.00)	<0.001
Glomerular filtration rate (mL/min)	111.00 (72.00–158.00)	129.00 (88.00–179.00)	109.00 (71.00–154.00)	<0.001
High-density lipoprotein (mmol/L)	1.10 (0.92–1.29)	1.09 (0.91–1.25)	1.11 (0.92–1.29)	0.029
Low-density lipoprotein (mmol/L)	2.85 (2.17–3.53)	2.79 (2.11–3.39)	2.89 (2.18–3.54)	0.027
Total bilirubin (µmol/L)	10.30 (7.50–13.90)	10.90 (7.80–14.15)	10.30 (7.40–13.80)	0.041
Direct bilirubin (µmol/L)	2.70 (1.90–3.70)	2.80 (2.00–3.80)	2.70 (1.90–3.70)	0.018
Albumin (g/L)	40.20 (37.10–42.80)	40.80 (38.20–43.10)	40.10 (37.00–42.70)	0.005
Lactate dehydrogenase (U/L)	161.00 (139.00–180.00)	158.00 (135.50–173.00)	161.00 (139.00–181.00)	0.001
AST/ALT	0.94 (0.74–1.24)	0.85 (0.64–1.10)	0.96 (0.75–1.25)	<0.001
Carbon dioxide combining power (mmol/L)	23.83 (22.20–25.40)	23.52 (22.00–24.90)	23.83 (22.20–25.50)	0.014
Lipoprotein(a) (mg/L)	95.00 (28.00–133.23)	93.00 (25.00–126.84)	95.00 (28.00–133.23)	0.026
Urea (mmol/L)	5.60 (4.40–7.02)	5.23 (4.06–6.40)	5.67 (4.50–7.10)	<0.001
Serum creatinine (µmol/L)	72.00 (57.00–89.00)	69.00 (55.00–83.00)	72.00 (58.00–89.25)	0.004
Haemoglobin (g/L)	130.00 (115.00–144.00)	136.00 (119.50–149.00)	130.00 (114.00–143.00)	<0.001
C-Peptide (ng/ml)	2.15 (1.45–2.77)	2.26 (1.68–3.05)	2.13 (1.43–2.75)	0.001
White cell count (10^9/L)	6.83 (5.65–8.19)	7.08 (5.79–8.45)	6.80 (5.63–8.17)	0.013
Red cell count (10^{12}/L)	4.49 (3.97–4.94)	4.68 (4.18–5.10)	4.46 (3.95–4.92)	<0.001
Absolute lymphocyte count (10^9/L)	1.95 (1.51–2.44)	2.08 (1.62–2.65)	1.93 (1.50–2.41)	<0.001
Neutrophil to lymphocyte ratio	2.03 (1.48–2.85)	1.91 (1.42–2.81)	2.05 (1.49–2.86)	0.027
Absolute eosinophil count (10^9/L)	0.16 (0.10–0.25)	0.14 (0.09–0.22)	0.16 (0.10–0.26)	0.014
Urinary microalbumin (mg/L)	22.20 (8.20–121.39)	17.00 (7.85–85.15)	22.80 (8.30–121.39)	0.007
Urinary microalbumin/creatinine ratio (mg/g)	27.20 (10.20–202.38)	19.00 (9.00–147.12)	28.90 (10.38–202.38)	<0.001
24-hour urinary protein (mg/24 hours)	111.00 (39.10–488.52)	88.90 (34.70–342.50)	114.00 (40.22–488.52)	0.001
Urinary creatinine (µmol/L)	7462.00 (4809.00–9924.00)	8484.66 (5478.50–11 213.00)	7320.50 (4726.75–9762.50)	<0.001
Free triiodothyronine (pg/mL)	2.91 (2.68–3.09)	3.02 (2.69–3.15)	2.91 (2.68–3.07)	<0.001
C reactive protein (mg/L)	10.40 (1.40–11.99)	7.90 (1.10–13.50)	11.20 (1.40–11.99)	0.007
Peripheral vascular disease	2608 (78.01)	208 (54.59)	2400 (81.03)	<0.001
Diabetic retinopathy	498 (14.90)	32 (8.40)	466 (15.73)	<0.001
Diabetic foot	541 (16.18)	39 (10.24)	502 (16.95)	0.001
Hypertension	1624 (48.58)	139 (36.48)	1485 (50.14)	<0.001
Fatty liver	1155 (34.55)	164 (43.04)	991 (33.46)	<0.001
Diabetic nephropathy	1138 (34.04)	103 (27.03)	1035 (34.94)	0.002

Continued

Table 1 Continued

Characteristics	Total (n=3343)	Non-DPN (n=381)	DPN (n=2962)	P value
Coronary atherosclerotic heart disease	311 (9.30)	23 (6.04)	288 (9.72)	0.020
Hypertensive heart disease	478 (14.30)	36 (9.45)	442 (14.92)	0.004
Cerebral infarction	1214 (36.31)	93 (24.41)	1121 (37.85)	<0.001

The continuous variables were shown as median (IQR) after the normality distribution test. The categorical variables were expressed as count (%).
AST/ALT, Aspartate Aminotransferase/Alanine Aminotransferase; BMI, body mass index; DPN, diabetic peripheral neuropathy.

SHAP values above zero indicating an increased risk of DPN, and higher values indicating a greater risk. Each patient's contribution to the model, through their individual feature values, is represented by dots, which are colour-coded, with red indicating higher and blue indicating lower values. Notably, figure 4 (left) identifies CRP as the most pivotal feature influencing DPN, exhibiting a complex and nonlinear relationship with DPN risk. Age exerts a strong positive impact, with increased age correlating with elevated risk for DPN. Conversely, GGT presents a negative association, where higher GGT values correspond to a reduced risk of DPN. Figure 4 (right) highlights the average impact of each feature on the model's outputs, demonstrating significant contributions from CRP, TBA, GGT and age, while factors like FT3, BMI and NLR exhibit lesser impacts.

The SHAP dependence plot can be used to understand how single features affect the output of the XGBoost model. Figure 5 (left) illustrates the dependency plot of age. The x-axis represents the raw values of age, while the y-axis indicates the corresponding SHAP values. When the SHAP value for a specific feature exceeds zero, it suggests an increased risk of DPN. As shown in the plot, for ages below 50, most SHAP values are below zero. Between 50 and 70 years old, all SHAP values exceed zero, and SHAP values increase as age increases. After the age of 70, the SHAP values tend to decrease with age. Figure 5 (right) displays the feature dependency between age and LDH. As age increases, more red dots appear on the plot, suggesting a potential positive correlation between LDH and age.

Model application

To further understand the contribution of features to individual patient predictions and to demonstrate the clinical application of the XGBoost model, we generated force plots for two randomly selected patients from the test set. In the force plot, red areas indicate features that increase the probability of DPN, while blue areas represent features that decrease this probability. The length of the arrows reflects the magnitude of each feature's impact on the prediction, with longer arrows indicating a greater effect. *f(x)* represents the comprehensive SHAP value of each patient, while the base value corresponds

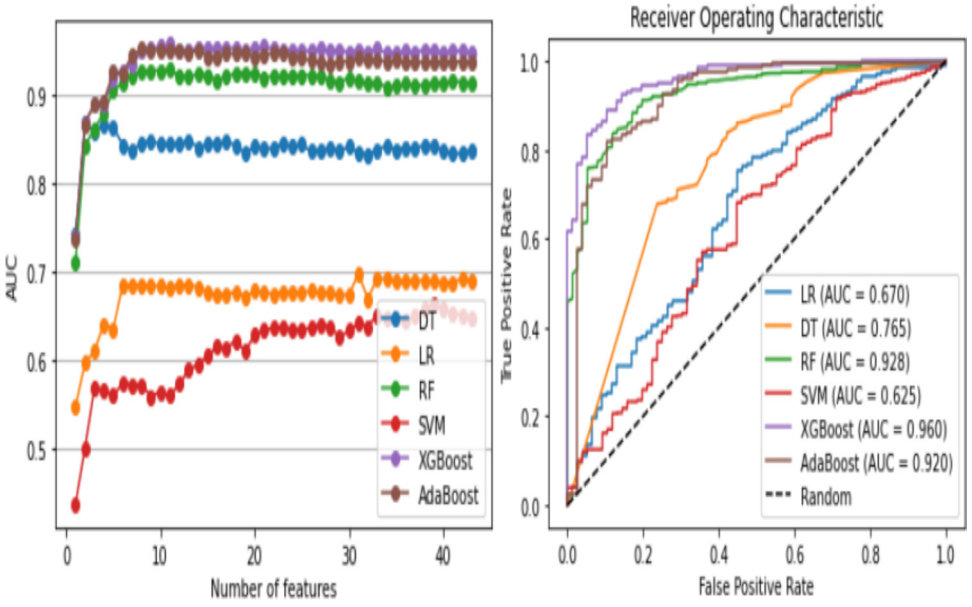

Figure 2 Change of AUC based on number of features (left) and ROC curve for different models on test set (right). AdaBoost, adaptive boosting; AUC, area under the curve; DT, decision tree; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; XGBoost, extreme gradient boosting.

Table 2 Performance comparison of different ML models

Model	AUC	G-mean	Accuracy	Precision	Recall	F1-score	FP	FN
LR	0.670	0.620	0.655	0.925	0.664	0.773	0.421	0.336
DT	0.765	0.693	0.771	0.940	0.793	0.860	0.395	0.207
RF	0.928	0.804	0.915	0.959	0.944	0.952	0.316	0.056
SVM	0.625	0.591	0.692	0.916	0.718	0.805	0.513	0.282
XGBoost	**0.960**	**0.850**	**0.927**	**0.969**	**0.948**	**0.958**	**0.237**	**0.052**
Adaboost	0.920	0.829	0.885	0.967	0.900	0.933	0.237	0.099

Bold values indicate the best results for each evaluation metric.
AdaBoost, adaptive boosting; AUC, area under the curve; DT, decision tree; FN, false negative rate; FP, false positive rate; LR, logistic regression; ML, machine learning; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

to the average SHAP value across all patients. If $f(x)$ exceeds the base value, the model predicts the patient has DPN. Figure 6 (above) shows that a patient with DPN was accurately predicted to have DPN by the model, with contributing factors including older age, lower GGT level and higher TBA level. Figure 6 (bottom) illustrates that the model correctly predicted a non-DPN patient not to have DPN, with the predicted reasons including higher GGT level and lower TBA level. These findings suggest that GGT has a negative impact on the risk of DPN, whereas TBA has a positive impact on DPN risk. Overall, the XGBoost model demonstrates strong discriminative power in distinguishing DPN from non-DPN patients, while the SHAP method can provide personalised insights into risk factors and probabilities based on each patient's individual characteristics.

DISCUSSION

DPN is a common and serious complication of diabetes, resulting from chronic hyperglycaemia. It significantly increases the risk of foot ulcers and lower limb amputations, which result in high rates of disability and mortality. Despite its serious impact, DPN is often overlooked in its early stages due to the non-specific symptoms and the limitations of nerve conduction studies in detecting small fibre lesions. Although measuring intraepidermal nerve fibre density from lower-limb skin biopsy is considered the gold standard for diagnosing small fibre neuropathy, its invasiveness makes it unsuitable for routine screening.[3] Therefore, it is crucial to develop a DPN risk prediction model to identify high-risk individuals for further targeted diagnosis.

Previous studies have used ML methods to establish prediction models for DPN. Yu *et al*[23] established a prediction model for painful DPN via LASSO regression; however, the specificity of their study population restricts the applicability of their findings to the general DPN patient population. Jiang *et al*[2] constructed a novel ML-based multifeatured Chinese-Western medicine-integrated prediction model for DPN using clinical features of traditional Chinese medicine indicators. However, the specificity of these indicators restricts their generalisability. Some studies have selected only one ML method,[22 23] thereby preventing a comparison with other methods and potentially reducing prediction accuracy. Lian *et al*[9] built six ML models and selected the optimal one, yet the limited sample size constrained the model's performance. Although Haque *et al*[24] applied various ML algorithms to large datasets and achieved excellent predictive performance, the 'black-box' nature of ML models limits their clinical application.[26 27] Therefore, it would be beneficial to construct a more interpretable and high-performing DPN prediction model using various ML methods based on a large dataset.

In this study, we developed six ML prediction models based on a large, high-dimensional dataset. Techniques such as RFE and SMOTE were employed to enhance predictive performance. The XGBoost model exhibited the highest performance with an AUC of 0.960 and recall of 0.948 on the independent test set. These metrics indicate that the model has good reliability. The significance of this study is that all the indicators used for model construction were extracted from real-world clinical data. Therefore, the model established can be used in a clinical setting to help clinicians identify whether a patient with T2DM is at high risk of DPN, even in the absence of obvious clinical symptoms or signs. The use of the SHAP methodology enables physicians to visualise the relationships between indicators and the DPN prediction model. The application of SHAP in individual predictions allows for the identification of specific influencing factors for each patient, thereby aiding clinicians in developing personalised prevention and treatment strategies.

The prevalence of DPN among patients with T2DM in this study was 88.6%, which was significantly higher than the previously reported range of 38.22%–67.6% in the China region.[32 45 46] The discrepancy may be attributed to several factors. First, different studies may have used varying diagnostic methods. Different methods may have different sensitivities and specificities in detecting DPN, which leads to different reported prevalences of DPN. Second, geographical differences could contribute to this variation. To date, there are no studies on the prevalence of DPN in Hainan. Hainan is a unique geographical

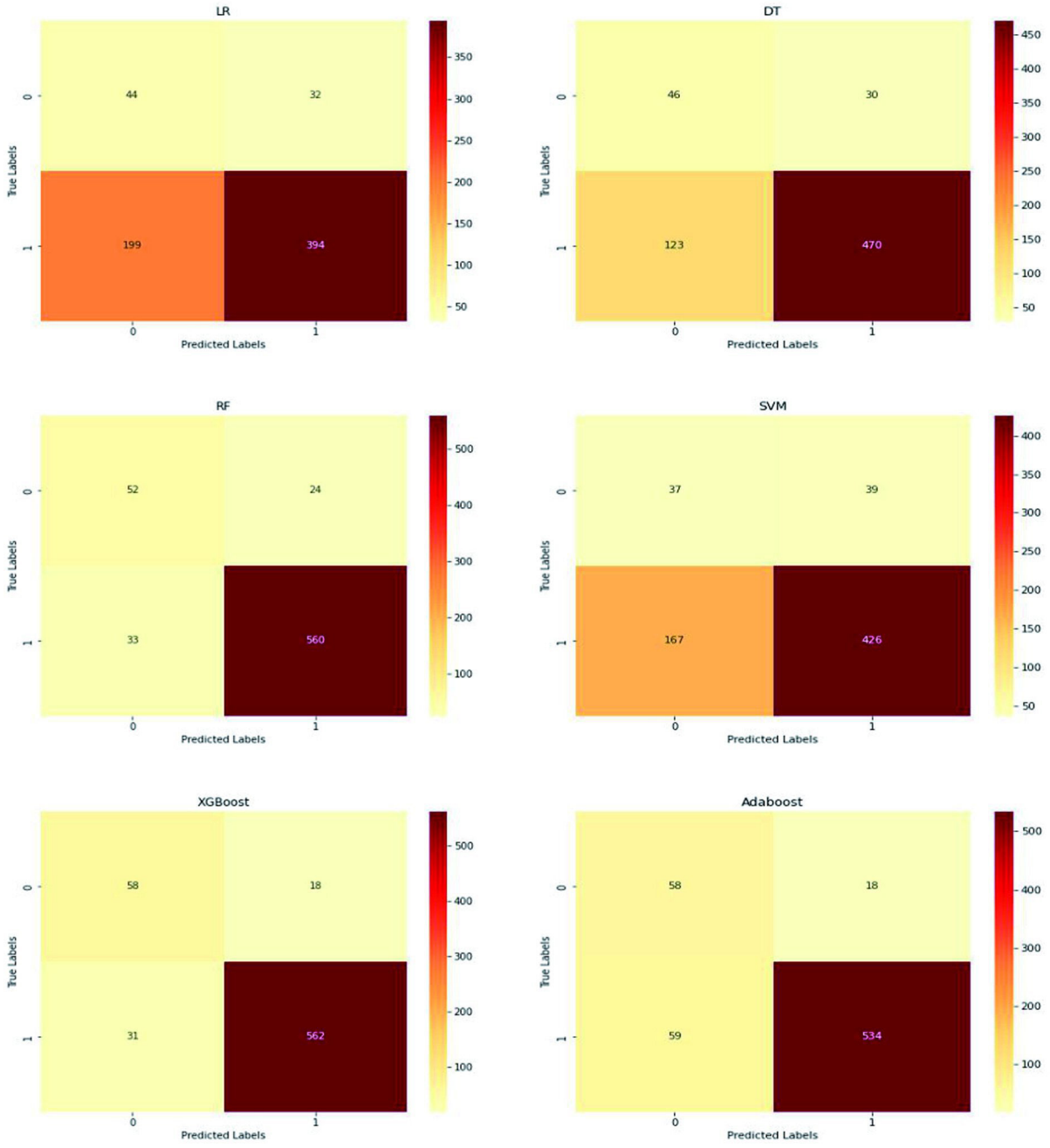

Figure 3 Confusion matrix of six machine learning models on the test set. AdaBoost, adaptive boosting; DT, decision tree; LR, logistic regression; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

region of China with a distinctive climate, diet and lifestyle. Hainan's tropical climate may lead to differences in physical activity levels, as people may be less likely to engage in outdoor activities during the hot and humid months. Reduced physical activity may lead to a higher prevalence of obesity and insulin resistance, which are known risk factors for DPN.[47 48] Third, our study population was recruited from a large comprehensive tertiary hospital in Hainan Province. Tertiary hospitals typically attract patients with more complex and severe medical conditions. The hospitalised patients in our study may have a longer duration of T2DM, poorer glycaemic control and more comorbidities compared with the general T2DM patient. Finally, the median age of our study participants was relatively high at 60 years, and age is a well-known risk factor for DPN.

Both the XGBoost model and the RF model identified CRP as the most important predictor variable. CRP is a common inflammatory marker in clinical practice. Albers and Pop-Busui[49] highlighted that the role of inflammatory response in the development of T2DM and its complications has attracted significant attention. CRP, induced by interleukin-6, is one of the sensitive inflammatory markers that can activate the nuclear factor κB pathway, inducing the production of various inflammatory mediators. It regulates cell apoptosis and promotes

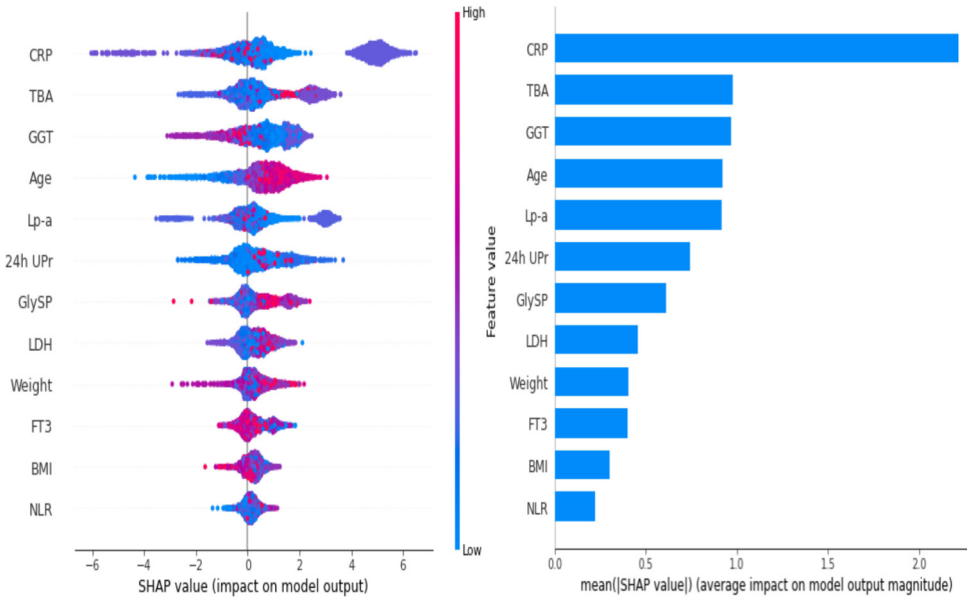

Figure 4 SHAP summary plot for the 12 clinical features contributing to the XGBoost model. BMI, body mass index; CRP, C reactive protein; GGT, gamma-glutamyl transpeptidase; LDH, lactate dehydrogenase; NLR, neutrophil-to-lymphocyte ratio; SHAP, Shapley Additive Explanation; TBA, total bile acid; XGBoost, extreme gradient boosting.

the occurrence of DPN. Several previous studies have also demonstrated that CRP levels are significantly elevated in T2DM patients with DPN, suggesting that CRP could serve as a useful prognostic biomarker to identify those with diabetes who are at risk of DPN.[25 50]

Age has been identified as a strong independent risk factor for DPN in previous studies.[10 32 51 52] As individuals age, some immune functions related to DPN may be further disrupted, contributing to an increased incidence of DPN over time.[53] T2DM is a chronic progressive disease, and as the disease advances and patient age increases, pancreatic β-cells experience a progressive decline in function, leading to T2DM-related complications. Meanwhile, as the disease progresses, pathological changes such as thickening of the basement membrane of microvessels in the body will occur, potentially leading to microvascular thrombosis. This can result in peripheral nerve necrosis and degeneration, ultimately contributing to the development of DPN.[54] Our study also confirms that age is an important risk factor for DPN in T2DM, aligning with previous research findings.

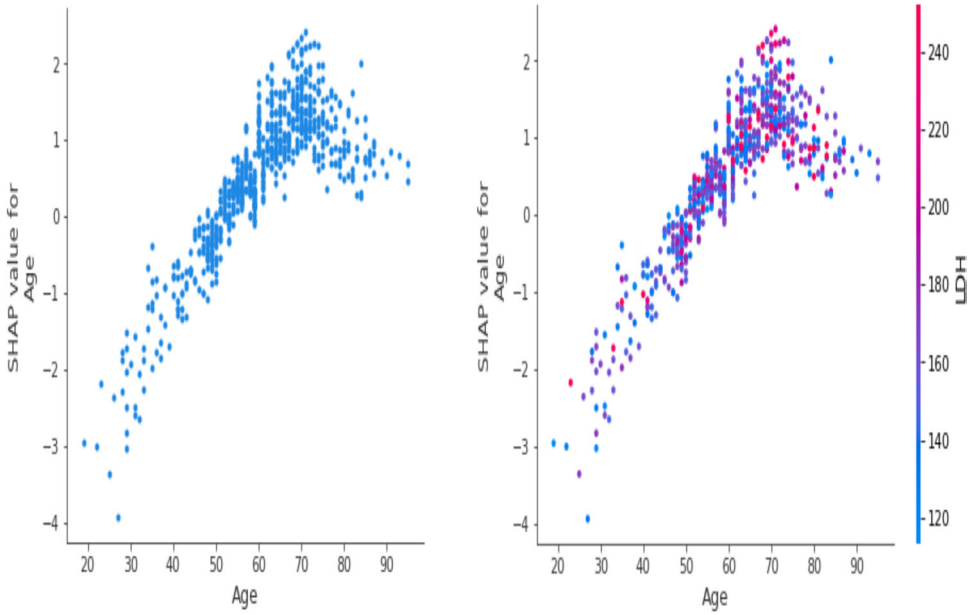

Figure 5 SHAP dependence plot of the XGBoost model. Feature dependency of age (left) and the mutual characteristic dependence between age and LDH (right). LDH, lactate dehydrogenase; SHAP, Shapley Additive Explanation; XGBoost, extreme gradient boosting.

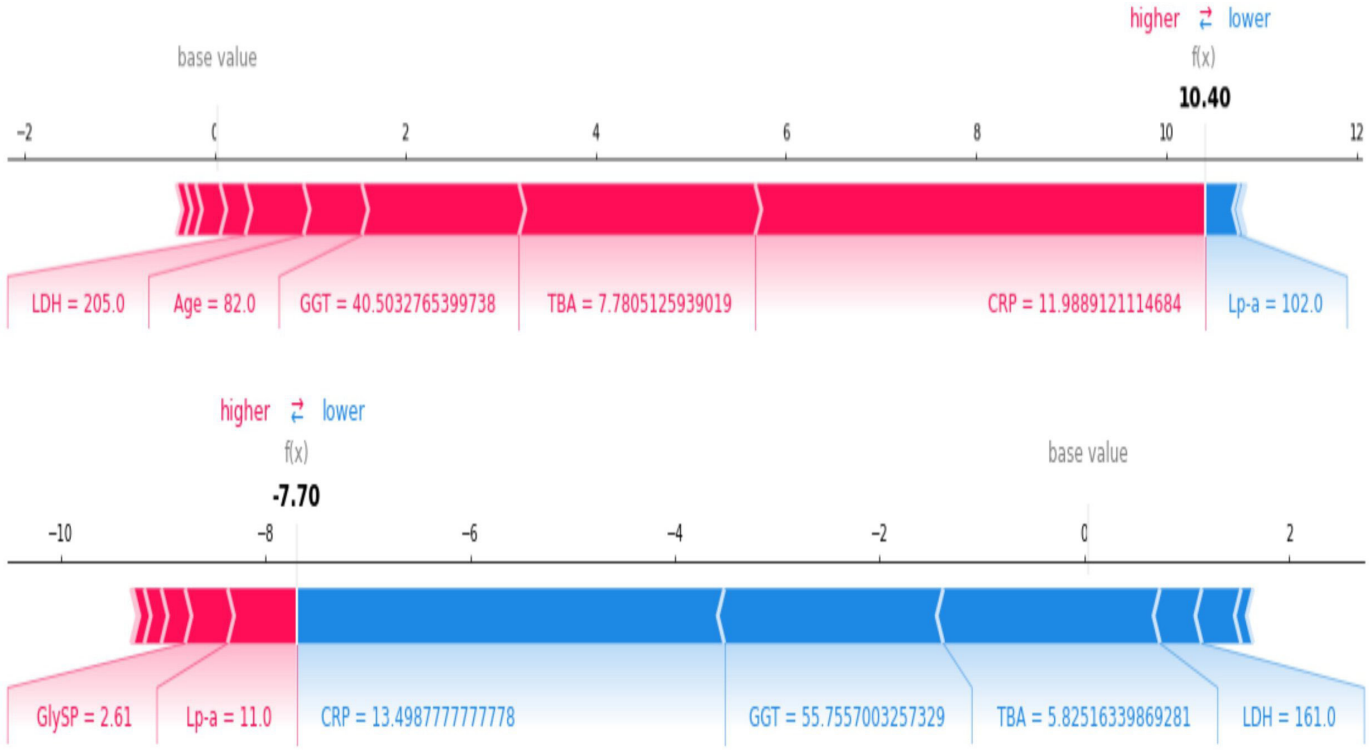

Figure 6 SHAP force plot for a DPN patient (above) and SHAP force plot for a non-DPN patient (below). CRP, C reactive protein; DPN, diabetic peripheral neuropathy; GGT, gamma-glutamyl transpeptidase; LDH, lactate dehydrogenase; SHAP, Shapley Additive Explanation; TBA, total bile acid.

Our study has identified several potential risk factors for DPN, such as 24-hour UPr and Lp-a, which are aligned with previous research.[9 31] In addition to these findings, our research has revealed novel insights that diverge from earlier studies. Specifically, we observed significantly elevated levels of TBA in DPN patients compared with non-DPN patients, suggesting that TBA may be a potential risk factor for DPN. However, our findings contrast with those of Yan *et al*,[55] who reported that lower levels of TBA might be associated with the development of DPN in individuals with T2DM. Given the limited number of existing studies on the relationship between TBA and DPN, and the cross-sectional nature of the available research, the exact causal relationship and underlying mechanisms remain unclear. Subsequent researchers can conduct longitudinal cohort studies to further elucidate the relationship between TBA and DPN. Similarly, El Boghdady and Badr[56] reported elevated levels of GGT in DPN patients and proposed that measuring serum GGT could aid in the early detection of DPN in diabetic populations. This finding is inconsistent with our results, highlighting the need for additional mechanistic studies to clarify the relationship between GGT and DPN, as current research on this topic is scarce.

Strengths and limitations

This retrospective study successfully developed and internally validated ML-based prediction models for DPN in T2DM. To enhance the reliability of the model, we implemented several strategies. First, we employed an RF-based RFE algorithm to identify the most influential feature subset. Additionally, we applied SMOTE to address data imbalance, thereby improving model performance. The XGBoost model was selected as the most effective model with high recall and accuracy on the original unbalanced test set, demonstrating its reliability. Furthermore, we employed the SHAP method to increase clinical utility by elucidating the relationship between feature and model predictions, thereby assisting clinicians in devising personalised prevention and treatment plans.

However, the study has several limitations. First, the data were retrospectively collected from electronic health records that lacked follow-up information, and the causal relationship between potential risk factors and DPN could not be verified. However, the findings can guide researchers in further investigations of the aetiology and underlying mechanisms of DPN. Second, the study was conducted at a single centre in Hainan and included only hospitalised T2DM patients. Although the model was validated using an independent test set, external validation in diverse T2DM populations is essential to confirm its applicability in other regions and settings. Cross-validation across multiple centres and the inclusion of ethnically, geographically and clinically diverse patient populations would significantly enhance the model's robustness and generalisability. Third, mean imputation could introduce bias. Subsequent refinement of the prediction model could attempt to use multiple filling methods and compare their effects, or use cross-validation to assess the impact of filling results on model

performance. Additionally, due to data limitations, we did not include or analyse patients' subjective descriptions, such as family history, duration of diabetes,[7 32] and smoking habits,[8] which have been identified as important factors in predicting the risk of DPN. Future research should integrate these subjective descriptors to refine the prediction models, thereby providing a more comprehensive and accurate model to predict DPN.

CONCLUSIONS

In this study, we developed several ML models to predict the risk of DPN in patients with T2DM. Among these models, XGBoost demonstrated the best performance. Its high AUC, accuracy, precision, recall and F1-score in the independent test set indicate that the model is reliable. Furthermore, the SHAP method facilitates the visualisation of the relationship between influencing factors and model outputs, enhancing the interpretability of the model. This interpretability is crucial for increasing physicians' trust in the prediction model. The application of this method in individual risk prediction and feature dependency analysis can help physicians identify high-risk patients and develop targeted preventive and treatment measures in resource-limited settings. Our team will address the existing issues, and future research will focus on conducting multicentre studies and integrating patients' subjective descriptions to enhance the model's predictive ability and generalisability. Additionally, to facilitate the use of the developed predictive models in clinical settings, a web interface should be developed to enable clinicians to input patients' feature values and press the 'Calculate' button to obtain the prediction result and SHAP analyses without needing to understand the complex principles behind the ML models.

Contributors The study conception and design were developed by LL, LC, XW, and FJ. Material preparation and data collection were by FJ. Data extraction was by BB. Data analysis was performed by LL, LZ and MG. The predictive model was established by LL. The first draft of the manuscript was written by LL. All authors commented on previous versions of the manuscript and all authors read and approved the final manuscript. LL is responsible for the overall content as guarantor. I have used AI for article polishing.

Funding This study was funded by the Natural Science Foundation of Hainan Province (Nos. 821QN0895 and 821MS044).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Ethics Committee of Hainan Medical University (Approval number: HYLL-2021-388). The study was conducted in accordance with the local legislation and institutional requirements. Since this is a retrospective study based on deidentified hospital information system data, informed consents were waived.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD

Lianhua Liu http://orcid.org/0009-0000-7884-6901

REFERENCES

1 Sun H, Saeedi P, Karuranga S, *et al*. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022;183:S0168-8227(21)00478-2.
2 Jiang A, Li J, Wang L, *et al*. Multi-feature, Chinese-Western medicine-integrated prediction model for diabetic peripheral neuropathy based on machine learning and SHAP. *Diabetes Metab Res Rev* 2024;40:e3801.
3 Selvarajah D, Kar D, Khunti K, *et al*. Diabetic peripheral neuropathy: advances in diagnosis and strategies for screening and early intervention. *Lancet Diabetes Endocrinol* 2019;7:938–48.
4 Pop-Busui R, Boulton AJM, Feldman EL, *et al*. Diabetic Neuropathy: A Position Statement by the American Diabetes Association. *Diabetes Care* 2017;40:136–54.
5 Zhang W, Chen L, Lou M. Association of Elevated Serum Uric Acid with Nerve Conduction Function and Peripheral Neuropathy Stratified by Gender and Age in Type 2 Diabetes Patients. *Brain Sci* 2022;12:1704.
6 Sen CK, Roy S, Khanna S. Diabetic Peripheral Neuropathy Associated with Foot Ulcer: One of a Kind. *Antioxid Redox Signal* 2023.
7 Gao L, Qin J, Chen Y, *et al*. Risk Factors for Subclinical Diabetic Peripheral Neuropathy in Type 2 Diabetes Mellitus. *Diabetes Metab Syndr Obes* 2024;17:417–26.
8 Liu X, Chen D, Fu H, *et al*. Development and validation of a risk prediction model for early diabetic peripheral neuropathy based on a systematic review and meta-analysis. *Front Public Health* 2023;11:1128069.
9 Lian X, Qi J, Yuan M, *et al*. Study on risk factors of diabetic peripheral neuropathy and establishment of a prediction model by machine learning. *BMC Med Inform Decis Mak* 2023;23:146.
10 Sloan G, Selvarajah D, Tesfaye S. Pathogenesis, diagnosis and clinical management of diabetic sensorimotor peripheral neuropathy. *Nat Rev Endocrinol* 2021;17:400–20.
11 Wu B, Niu Z, Hu F. Study on Risk Factors of Peripheral Neuropathy in Type 2 Diabetes Mellitus and Establishment of Prediction Model. *Diabetes Metab J* 2021;45:526–38.
12 Li Y, Li Y, Deng N, *et al*. Training and External Validation of a Predict Nomogram for Type 2 Diabetic Peripheral Neuropathy. *Diagnostics (Basel)* 2023;13:1265.
13 Alshammari NA, Alodhayani AA, Joy SS, *et al*. Evaluation of Risk Factors for Diabetic Peripheral Neuropathy Among Saudi Type 2 Diabetic Patients with Longer Duration of Diabetes. *Diabetes Metab Syndr Obes* 2022;15:3007–14.
14 Noorbakhsh-Sabet N, Zand R, Zhang Y, *et al*. Artificial Intelligence Transforms the Future of Health Care. *Am J Med* 2019;132:795–801.
15 Handelman GS, Kok HK, Chandra RV, *et al*. eDoctor: machine learning and the future of medicine. *J Intern Med* 2018;284:603–19.
16 Tan KR, Seng JJB, Kwan YH, *et al*. Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. *J Diabetes Sci Technol* 2023;17:474–89.
17 Ejiyi CJ, Qin Z, Amos J, *et al*. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthcare Analytics* 2023;3:100166.
18 He F, Ng Yin Ling C, Nusinovici S, *et al*. Development and External Validation of Machine Learning Models for Diabetic Microvascular Complications: Cross-Sectional Study With Metabolites. *J Med Internet Res* 2024;26:e41065.
19 Li W, Song Y, Chen K, *et al*. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *BMJ Open* 2021;11:e050989.

20 Allen A, Iqbal Z, Green-Saxena A, *et al*. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ Open Diabetes Res Care* 2022;10:e002560.
21 Baskozos G, Themistocleous AC, Hebert HL, *et al*. Classification of painful or painless diabetic peripheral neuropathy and identification of the most powerful predictors using machine learning models in large cross-sectional cohorts. *BMC Med Inform Decis Mak* 2022;22:144.
22 Wei Z, Wang X, Lu L, *et al*. Construction of an Early Risk Prediction Model for Type 2 Diabetic Peripheral Neuropathy Based on Random Forest. *Comput Inform Nurs* 2024;42:665–74.
23 Yu Z, Zhao S, Cao J, *et al*. Analysis of risk factors for painful diabetic peripheral neuropathy and construction of a prediction model based on Lasso regression. *Front Endocrinol (Lausanne)* 2024;15:1477570.
24 Haque F, Bin Ibne Reaz M, Chowdhury MEH, *et al*. Performance Analysis of Conventional Machine Learning Algorithms for Diabetic Sensorimotor Polyneuropathy Severity Classification. *Diagnostics (Basel)* 2021;11:801.
25 Allwright M, Karrasch JF, O'Brien JA, *et al*. Machine learning analysis of the UK Biobank reveals prognostic and diagnostic immune biomarkers for polyneuropathy and neuropathic pain in diabetes. *Diabetes Res Clin Pract* 2023;201:S0168-8227(23)00488-6:110725:.
26 The Lancet Respiratory M. Opening the black box of machine learning. *Lancet Respir Med* 2018;6.
27 Watson DS, Krutzinna J, Bruce IN, *et al*. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 2019;364:l886.
28 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Sys-tems 2017:4765–74*.
29 Hosseini Sarkhosh SM, Hemmatabadi M, Esteghamati A. Development and validation of a risk score for diabetic kidney disease prediction in type 2 diabetes patients: a machine learning approach. *J Endocrinol Invest* 2023;46:415–23.
30 Wang S, Dai Y, Shen J, *et al*. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci Rep* 2021;11:24039.
31 Chen T, Xiao S, Chen Z, *et al*. Risk factors for peripheral artery disease and diabetic peripheral neuropathy among patients with type 2 diabetes. *Diabetes Res Clin Pract* 2024;207:S0168-8227(23)00842-2:111079:.
32 Wang W, Ji Q, Ran X, *et al*. Prevalence and risk factors of diabetic peripheral neuropathy: A population-based cross-sectional study in China. *Diabetes Metab Res Rev* 2023;39:e3702.
33 Liu J, Yuan X, Liu J, *et al*. n.d. Risk Factors for Diabetic Peripheral Neuropathy, Peripheral Artery Disease, and Foot Deformity Among the Population With Diabetes in Beijing, China: A Multicenter, Cross-Sectional Study. *Front Endocrinol*13:824215.
34 Cheng Y, Cao W, Zhang J, *et al*. Determinants of Diabetic Peripheral Neuropathy and Their Clinical Significance: A Retrospective Cohort Study. *Front Endocrinol* 2022;13:934020.
35 Li J, Liu S, Hu Y, *et al*. Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. *J Med Internet Res* 2022;24:e38082.
36 Sun Y, Wong AKC, Kamel MS. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *Int J Patt Recogn Artif Intell* 2009;23:687–719.
37 Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. *Comput Biol Med* 2021;128:S0010-4825(20)30420-0:104089:.
38 Schober P, Vetter TR. Logistic Regression in Medical Research. *Anesthesia & Analgesia* 2021;132:365–6.
39 Hong W, Lu Y, Zhou X, *et al*. Usefulness of Random Forest Algorithm in Predicting Severe Acute Pancreatitis. *Front Cell Infect Microbiol* 2022;12:893294.
40 Valkenborg D, Rousseau A-J, Geubbelmans M, *et al*. Support vector machines. *Am J Orthod Dentofacial Orthop* 2023;164:754–7.
41 Li Q, Yang H, Wang P, *et al*. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *J Transl Med* 2022;20:177.
42 Baniasadi A, Rezaeirad S, Zare H, *et al*. Two-Step Imputation and AdaBoost-Based Classification for Early Prediction of Sepsis on Imbalanced Clinical Data. *Crit Care Med* 2021;49:e91–7.
43 Chadha A, Kaushik B. A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data. *New Gener Comput* 2022;40:889–914.
44 Lundberg SM, Erion G, Chen H, *et al*. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020;2:56–67.
45 Pan Q, Li Q, Deng W, *et al*. Prevalence of and Risk Factors for Peripheral Neuropathy in Chinese Patients With Diabetes: A Multicenter Cross-Sectional Study. *Front Endocrinol (Lausanne)* 2018;9:617.
46 Wang L, Gao P, Zhang M, *et al*. Prevalence and Ethnic Pattern of Diabetes and Prediabetes in China in 2013. *JAMA* 2017;317:2515–23.
47 Kamal A, Aleem S, Kamal A, *et al*. The Risk factors of diabetic neuropathy in type 2 diabetic patient in Services Hospital Lahore. *Esculap* 2023;19:72–7.
48 Basebaa A, Musiaan N, Mahross A. Prevalence and Risk Factors of Diabetic Peripheral Neuropathy: A cross-sectional study from Yemen. *Ain Shams Med J* 2024;75:201–13.
49 Albers JW, Pop-Busui R. Diabetic neuropathy: mechanisms, emerging treatments, and subtypes. *Curr Neurol Neurosci Rep* 2014;14:473.
50 Ge S, Xie J, Zheng L, *et al*. Associations of serum anti-ganglioside antibodies and inflammatory markers in diabetic peripheral neuropathy. *Diabetes Res Clin Pract* 2016;115:S0168-8227(16)30009-2:68–75:.
51 Mao F, Zhu X, Liu S, *et al*. Age as an Independent Risk Factor for Diabetic Peripheral Neuropathy in Chinese Patients with Type 2 Diabetes. *Aging Dis* 2019;10:592–600.
52 Li L, Chen J, Wang J, *et al*. Prevalence and risk factors of diabetic peripheral neuropathy in Type 2 diabetes mellitus patients with overweight/obese in Guangdong province, China. *Prim Care Diabetes* 2015;9:S1751-9918(14)00082-5:191–5:.
53 Hagen KM, Ousman SS. Aging and the immune response in diabetic peripheral neuropathy. *J Neuroimmunol* 2021;355:S0165-5728(21)00101-6:577574:.
54 Guo M, Li M, Lv S, *et al*. Influence factors of subclinical diabetic peripheral neuropathy in patients with type 2 diabetes mellitus. *Chin J Prev Contr Chron Dis* 2023;31:289–93.
55 Yan P, Wan Q, Zhang Z, *et al*. Decreased Physiological Serum Total Bile Acid Concentrations in Patients with Type 2 Diabetic Peripheral Neuropathy. *Diabetes Metab Syndr Obes* 2021;14:2883–92.
56 El Boghdady NA, Badr GA. Evaluation of oxidative stress markers and vascular risk factors in patients with diabetic peripheral neuropathy. *Cell Biochem Funct* 2012;30:328–34.