Clinical and Experimental Medicine (2025) 25:151
https://doi.org/10.1007/s10238-025-01697-w

REVIEW

Machine learning for the prediction of diabetes-related amputation: a systematic review and meta-analysis of diagnostic test accuracy

Zhigang Chen[1] · Xinliang Liu[2] · Simeng Li[1] · Zhenheng Wu[3] · Haifen Tan[4] · Fuqian Yu[5] · Dongmei Wang[1] · Yawen Bo[6]

Received: 14 March 2025 / Accepted: 14 April 2025

Abstract

Although machine learning is frequently used in medicine for predictive purposes, its accuracy in diabetes-related amputation (DRA) remains unclear. From establishing the database until December 2024, we conducted a comprehensive search of PubMed, Web of Science (WoS), Embase, Scopus, Cochrane Library, Wanfang, and the China National Knowledge Index (CNKI). The pooled sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), diagnostic odds ratio (DOR), area under the curve (AUC), and Fagan plot analysis were used to assess the overall test performance of machine learning. Moreover, subgroup analysis and meta-regression were performed to search for possible sources of heterogeneity. Finally, sensitivity analysis and Deeks' funnel plot asymmetry test were used to evaluate the stability and publication bias, respectively. In the end, seven publications were included in this meta-analysis. The overall pooled diagnostic data were as follows: sensitivity, 0.72 (95% CI 0.69–0.75); specificity, 0.89 (95% CI 0.84–0.93); PLR, 3.62 (95% CI 3.36–3.89); NLR, 0.32 (95% CI 0.30–0.35); DOR, 13.55 (95% CI 11.72–15.67). The AUC was 0.81 (95% CI 0.77–0.84). The Fagan plot analysis showed that the positive post-test probability is 62% and the negative post-test probability is 7%. Subgroup analysis and meta-regression showed that both the level of bias and the year of publication were sources of heterogeneity in sensitivity and specificity. Sensitivity analysis confirmed the robustness of the results after excluding three outlier studies. The Deeks' funnel plot suggests that publication bias has no statistical significance ($P > 0.05$). In summary, our results suggest the moderate accuracy of machine learning in predicting DRA.

Keywords Machine learning · Diabetes-related amputation (DRA) · Accuracy · AUC · Prediction

Abbreviations

DRA	Diabetes-related amputation
WoS	Web of Science
CNKI	China National Knowledge Infrastructure
PLR	Positive Likelihood Ratio
NLR	Negative Likelihood Ratio
DOR	Diagnostic Odds Ratio
AUC	Area Under the Curve

Zhigang Chen, Xinliang Liu, Simeng Li and Zhenheng Wu have contributed equally to this work.

✉ Dongmei Wang
dongmeiwang0526@163.com

✉ Yawen Bo
byw780820@sina.com

[1] Department of Gastrointestinal Surgery, Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University, The Third Affiliated Hospital of Nanjing Medical University, Changzhou Medical Center, Nanjing Medical University, No. 68 Gehu Road, Wujin District, Changzhou City 213000, Jiangsu, China

[2] Department of Radiation Oncology, Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University, The Third Affiliated Hospital of Nanjing Medical University, Changzhou Medical Center, Nanjing Medical University, Changzhou 213000, Jiangsu, China

[3] Department of Hepatobiliary Surgery, Fujian Medical University Union Hospital, Fujian Medical University, Fuzhou 350000, China

[4] Department of Oral Surgery, Affiliated Hospital of Guangdong Medical University, Zhanjiang 524001, China

[5] Gastroenterology Department, The Second Affiliated Hospital of Anhui Medical University, Anhui Medical University, Hefei 230000, China

[6] Department of Endocrinology, Changzhou Second People's Hospital, Affiliated Changzhou No. 2 People's Hospital of Nanjing Medical University, The Third Affiliated Hospital of Nanjing Medical University, Changzhou Medical Center, Nanjing Medical University, No. 29 Xinglong Road, ChangzhouJiangsu 213000, China

Published online: 10 May 2025

CI	Confidence Interval
SROC	Summary Receiver Operating Characteristic
PRISMA-DTA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Diagnostic Test Accuracy
PROSPERO	Prospective Register of Systematic Reviews
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2
GRADE	Grading of Recommendations Assessment, Development and Evaluation
I^2	I-squared
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
AI	Artificial Intelligence
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
EHR	Electronic Health Records
HCC	Hepatocellular Carcinoma
AnF	Any Liver Fibrosis
RF	Random Forest
DT	Decision Tree
LR	Likelihood Ratio
XGBoost	Extreme Gradient Boosting

Introduction

Diabetes-related amputation (DRA) is a severe complication of diabetes mellitus, affecting a significant proportion of the diabetic population [1]. Approximately 85% of all non-traumatic amputations occur in patients with diabetes, and at least 80% of those are preceded by active foot lesions [2]. DRA and a high risk of amputation and mortality can impact patients' quality of life, life roles, and body image, as well as the financial burden placed on patients and their families [3–5]. In most cases, many treatments are not enough to cure diabetes and lead to amputation [6]. Therefore, it is essential to predict the risk of amputation associated with diabetes as early as possible and to intervene in a timely manner to mitigate its harmful consequences on patients' mental health and quality of life.

Machine learning and big data use in medicine have increased in recent years [7–9]. Because of the strong correlation between the visual aspects of human diseases and their diagnostic intuitiveness, the utilization of machine learning in diagnosing such ailments has witnessed rapid advancements [10]. Certain machine-learning products have now entered clinical implementation [11, 12].

Machine-learning techniques have gained a lot of interest in recent years for DRA monitoring and diagnosis in patients with neuropathic diabetes [13–18]. For instance, a Chinese study developed a machine learning system (XGBoost for DRA) for predicting DRA among individuals, achieving a accuracy of up to 80% [15]. Moreover, another study developed a tree machine learning system, and the accuracy rate increased to 81.4% [16]. In another study, a deep learning model was utilized in the risk assessment and detection of DRA [17]. Although research on machine learning involvement in DRA prediction is expanding, the accuracy of machine learning varies significantly across studies, which is attributable to differing algorithmic approaches [18].

Considering the aforementioned factors, along with the absence of systematic reviews and meta-analyses assessing the accuracy of machine learning-specific diagnoses for DRA, this study seeks to contribute evidence in this area. DRA significantly impact morbidity, mortality, and quality of life among users of public health services, with early identification of high-risk patients is critical for timely intervention. Furthermore, diagnostic methods such as Computed tomography angiography (CTA) and magnetic resonance imaging (MRI)—which are non-invasive, readily accessible, and purportedly effective in accurately identifying DRA lesions—are available within public health systems. Thus, undertaking a systematic review and meta-analysis to assess the predictive accuracy of machine learning models in forecasting DRA risks may yield significant insights into their efficacy in predicting DRA.

Materials and methods

Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) guidelines suggested reporting items were followed in the course of this study [19, 20]. See Supplemental Material 1 for a detailed list of PRISMA-DTA. The protocol was registered on the International Prospective Register of Ongoing Systematic Reviews (PROSPERO Number: CRD42024570834 https://www.crd.york.ac.uk/prospero/). This study was not subject to institutional review board oversight, and patient informed consent requirements because it was based solely on published trials.

Search strategy

Two independent researchers carried out a thorough literature search of the databases from inception to December 2024, including PubMed, Web of Science (WoS), Embase, Scopus, Cochrane Library, Wanfang, and the China National

Knowledge Index (CNKI). The languages involved only English and Chinese.

Search terms included “AI”, “Artificial intelligence”, “Machine learning model”, “Machine learning”, “Deep learning”, “Model”, “Diabetes”, “Amputation”, “Diabetes-related amputation”, “DRA”, “Sensitivity”, “Specificity”, “AUC”, “Area under the curve”, “True positive”, “TP”, “False positive”, “FP”, “F1-score”. Detailed search strategies are shown in Supplementary Material 2. An extended search of relevant reviews and references of included articles was conducted. The search method is adjusted according to the characteristics of the database and the search results by combining subject words and free words. Search results were imported into NoteExpress 3.4 (Beijing Aegean Hailezhi Technology Co.), and duplicates were removed.

Inclusion criteria and exclusion criteria

Inclusion criteria: (1) Published articles evaluating the value of machine learning in predicting the risk of DRA; (2) Cohort studies with diagnostic experiments; (3) The diagnostic criteria for DRA are clearly described in the literature; (4) Sensitivity, specificity, and optimal diagnostic thresholds are indicated in the literature or can be calculated.

Exclusion criteria: (1) Systematic review, review, and cases; (2) incomplete data, such as the necessary data such as the sensitivity and specificity of the item cannot be calculated; (3) Duplication of literature; (4) The full text is not available; (5) Challenged, or withdrawn literature.

Data extraction

Studies that fulfilled the specified criteria for inclusion and exclusion were meticulously assessed and, if required, approved or rejected. Subsequently, data pertaining to the diagnostic test under consideration, the index test under consideration, as well as the occurrences of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were extracted from the selected studies. The researchers contacted the authors of any paper that was found to have incomplete information and requested them to provide any missing details.

Quality evaluation

Publication quality was evaluated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) quality scoring standard [21]. The scale contains 14 key questions, each judged by a yes, no, or unclear answer. If the two researchers have differences in literature retrieval, screening, data extraction and literature quality evaluation, they would negotiate with the third researcher to resolve the differences.

Diagnostic performance and clinical value verification

A summary receiver operating characteristic curve (SROC) was drawn and AUC was calculated. Publication bias was evaluated using Deeks’ funnel plot asymmetry test, which is the recommended method for diagnostic test accuracy meta-analyses to assess the association between sample size and DOR [22]. A significance level of $P < 0.05$ was considered indicative of potential bias, as per the Cochrane Handbook guidelines [23]. Fagan chart was drawn to calculate the pre-test probability and post-test probability to evaluate the clinical value [24].

Bivariate model analysis

Threshold effects were assessed using Spearman’s correlation coefficient between sensitivity and (1-specificity), with $P < 0.05$ indicating significant threshold effect [25]. In addition, the bivariate random-effects model was employed to account for the inherent correlation between sensitivity and specificity, providing pooled estimates with 95% confidence regions [26].

Heterogeneity analysis

The statistical I^2 and Q tests were used to analyze the heterogeneity. An I^2 value > 50% was considered indicative of substantial heterogeneity, and > 75% reflected severe heterogeneity. If the heterogeneity was large, the random effect size model was used. If the heterogeneity is small, the fixed effect size model is used. Subgroup analysis and meta-regression analysis were performed to explore the sources of heterogeneity. Variables included were predefined as follows: sample size (≤ 1,000 vs. > 1,000), publication year (before 2020 vs. 2020–2024), machine learning algorithm type (RF vs. Others), and risk of bias level (QUADAS-2 score ≤ 2 vs. > 2).

Quality of evidence assessment

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) method was used to systematically evaluate the body of evidence [27]. The quality of evidence was independently evaluated by two researchers across the following dimensions, ultimately categorizing the evidence into four levels: high, medium, low, and very low. These dimensions included: (1) Risk of bias, assessed using the QUADAS-2 tool; (2) Inconsistency, determined by the heterogeneity test results (I^2 value); (3) Indirectness, evaluated by examining the relevance of the study population and intervention to the clinical question; (4) Imprecision, assessed based on the width of the confidence interval and the clinical decision threshold; and (5) Publication bias, analyzed using Deeks' funnel plot.

Statistical analysis

The statistical analysis was conducted using RevMan 5.4 software (RevMan 5.4, Review Manager, Version 5.4, The Cochrane Collaboration), meta-disc software, and Stata version 16.0 (StataCorp LP, USA) [28–30]. RevMan 5.4 software was used to generate quality evaluation plots. Stata version 16.0 was used to analyze SROC curve and sensitivity analysis, etc. meta-disc software was used to generate PLR, NLR, and the DOR. $P < 0.05$ was considered statistically significant.

Result

Search results

Following a comprehensive examination of the titles and abstracts of all 1785 articles, 1105 were excluded due to duplication, and 492 were dismissed for failing to meet the inclusion criteria. Subsequent to a full-text review, an additional 181 articles were excluded. Ultimately, seven studies satisfied the eligibility requirements and were included in the analysis. A detailed overview of the literature screening process is presented in Fig. 1.

Basic characteristics

Table 1 summarizes all included studies. Among the seven studies, the largest number came from China (3 in total), two from the United States, one from Spain and one from Turkey. All studies were retrospective cohort studies. A total of 105,928 patients were included in all studies, including 6220 in the Experiment group and 99,708 in the control group. The participants' ages ranged from 26 to 88. The highest proportion of males was 65%, while the lowest was 53.8%.

QUADAS-2 assessment

Supplementary Fig. 1 and Supplementary Table 1 display the QUADAS-2 quality assessment results. The majority of the publications that were part of the current meta-analysis satisfied the majority of the QUADAS-2 items, indicating that the general caliber of the studies that were included was moderate to high. Most studies scored three or above therefore considered to be of good quality. Two studies scored

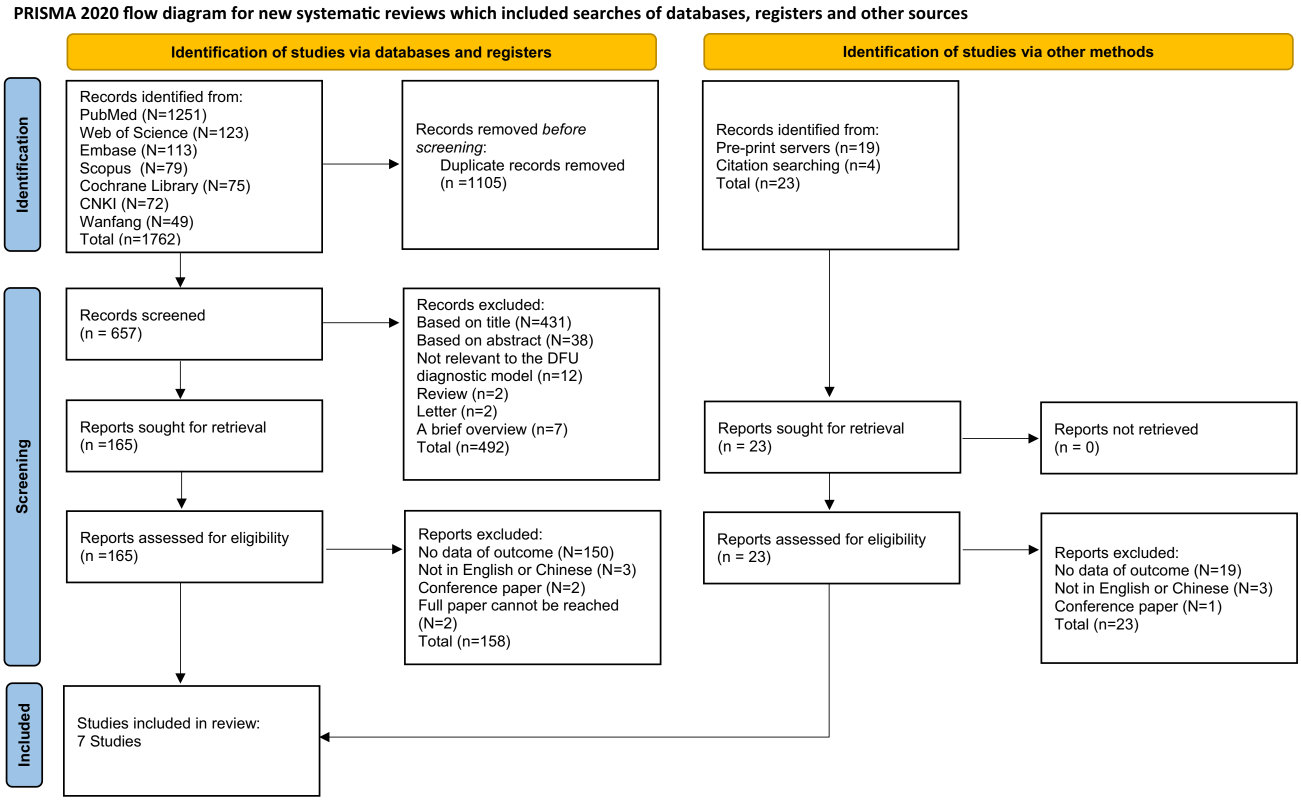
PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from:
PubMed (N=1251)
Web of Science (N=123)
Embase (N=113)
Scopus (N=79)
Cochrane Library (N=75)
CNKI (N=72)
Wanfang (N=49)
Total (n=1762)
Records removed before screening:
Duplicate records removed (n =1105)
Records identified from:
Pre-print servers (n=19)
Citation searching (n=4)
Total (n=23)
Records screened (n = 657)
Records excluded:
Based on title (N=431)
Based on abstract (N=38)
Not relevant to the DFU diagnostic model (n=12)
Review (n=2)
Letter (n=2)
A brief overview (n=7)
Total (n=492)
Reports sought for retrieval (n =165)
Reports sought for retrieval (n = 23)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n =165)
Reports excluded:
No data of outcome (N=150)
Not in English or Chinese (N=3)
Conference paper (N=2)
Full paper cannot be reached (N=2)
Total (n=158)
Reports assessed for eligibility (n = 23)
Reports excluded:
No data of outcome (N=19)
Not in English or Chinese (N=3)
Conference paper (N=1)
Total (n=23)
Studies included in review:
7 Studies

Fig. 1 Flow diagram of the meta-analysis

Table 1 Characteristics of included studies

Authors	Country	Research type	Machine learning type	Experiment	Age	Male (%)	Control	Age	Male(%)	Sensitivity/recall	Specificity	AUC	Accuracy	Precision	F1
Denizhan Demirkol et al. 2023	Turkey	Retrospective	RF-hyperparameter optimization BO hybrid approach	179	Not mentioned	Not mentioned	228	Not mentioned	Not mentioned	0.77	0.94	Not mentioned	0.85	0.91	0.83
Chenzhen Du et al. 2021	China	Retrospective	XGBoost	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0.67	0.86	0.86	0.8	0.827	0.74
Chenzhen Du et al. 2021	China	Retrospective	LR	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0.33	0.86	0.76	0.7	0.702	0.449
Chenzhen Du et al. 2021	China	Retrospective	SVM	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0	1	0.6	0.7	0	0
Chenzhen Du et al. 2021	China	Retrospective	RF	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0.67	0.86	0.67	0.8	0.827	0.74
Chenzhen Du et al. 2021	China	Retrospective	GBDT	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0.33	0.86	0.67	0.7	0.702	0.449
Chenzhen Du et al. 2021	China	Retrospective	ANN	3	Not mentioned	Not mentioned	7	Not mentioned	Not mentioned	0.33	0.86	0.71	0.7	0.702	0.449
Puguang Xie et al. 2022	China	Retrospective	Multi-class classification model-minor amputation	6	68.1 ± 10.4	60.6	100	66.0 ± 12.3	62.6	0.643	0.945	0.85	0.719	0.972	0.774

Table 1 (continued)

Authors	Country	Research type	Machine learning type	Experiment	Age	Male (%)	Control	Age	Male(%)	Sensitivity/recall	Specificity	AUC	Accuracy	Precision	F1
Puguang Xie et al. 2022	China	Retrospective	Multi-class classification model-major amputation	8	66.4 ± 12.7	55.3	100	66.0 ± 12.3	62.6	0.333	0.973	0.86	0.493	0.974	0.496
Puguang Xie et al. 2022	China	Retrospective	Multi-class classification model-overall	24	Not mentioned	Not mentioned	100	66.0 ± 12.3	62.6	0.871	0.744	0.9	0.839	0.911	0.89
Shiqi Wang et al. 2022	China	Retrospective	After oversampling-DT	86	26–88	Not mentioned	86	26–88	Not mentioned	0.616	0.967	0.813	0.744	0.828	0.707
Shiqi Wang et al. 2022	China	Retrospective	After oversampling-RF	86	26–88	Not mentioned	86	26–88	Not mentioned	0.756	0.958	0.857	0.797	0.823	0.788
Shiqi Wang et al. 2022	China	Retrospective	After oversampling-LR	86	26–88	Not mentioned	86	26–88	Not mentioned	0.64	0.906	0.739	0.64	0.64	0.64
Shiqi Wang et al. 2022	China	Retrospective	After oversampling-SVM	86	26–88	Not mentioned	86	26–88	Not mentioned	0.593	0.93	0.767	0.663	0.689	0.638
Shiqi Wang et al. 2022	China	Retrospective	After oversampling-XGBoost	86	26–88	Not mentioned	86	26–88	Not mentioned	0.767	0.964	0.881	0.814	0.846	0.805
Jose´ Barbera´n et al. 2010	Spain	Retrospective	Logistic regression model	26	70.57 ± 9.6	65	52	68.07 ± 10.8	53.8	0.962	0.788	0.93	0.846	0.694	0.806
Stavros Stefanopoulos et al. 2022	USA	Retrospective	CTREE-10 variables	5803	Not mentioned	Not mentioned	92,253	Not mentioned	Not mentioned	0.77	0.778	0.84	0.77	0.982	0.863

Table 1 (continued)

Authors	Country	Research type	Machine learning type	Experiment	Age	Male (%)	Control	Age	Male(%)	Sensitivity/recall	Specificity	AUC	Accuracy	Precision	F1
Stavros Stefanopoulos et al. 2022	USA	Retrospective	CTREE-5 variables	5803	Not mentioned	Not mentioned	92,253	Not mentioned	Not mentioned	0.762	0.794	0.84	0.764	0.983	0.859
Stavros Stefanopoulos et al. 2022	USA	Retrospective	RF	5803	Not mentioned	Not mentioned	92,253	Not mentioned	Not mentioned	0.757	0.798	0.83	0.759	0.983	0.856
Lanting Yang et al. 2021	USA	Retrospective	LASSO	39	Not mentioned	Not mentioned	6922	Not mentioned	Not mentioned	0.718	0.695	Not mentioned	0.707	0.013	0.026
Lanting Yang et al. 2021	USA	Retrospective	GBM	39	Not mentioned	Not mentioned	6922	Not mentioned	Not mentioned	0.718	0.695	Not mentioned	0.707	0.013	0.026

RF Random Forest Algorithm, *CART* classification and regression trees, *XGBoost* The extreme gradient boosting, *LR* logistic regression, *SVM* support vector machine, *ANN* artificial neural network, *GBDT* Gradient Boosting Decision Tree, *DT* Decision Tree, *DFU* Diabetic Foot Ulcer, *CTREE* Conditional Inference Tree, *LASSO* Least Absolute Shrinkage and Selection Operator, *GBM* Gradient Boosting Machine two, indicating poor quality, but after a detailed review of the paper and its evidence, the papers were included in the final analysis. This is because although blinding was not mentioned, the selection process for participants in the continuous enrollment of patients was reasonable, and there were no concerns about its applicability. This paper meets the selection criteria for our study, otherwise consistent with other studies included.

Machine learning can accurately predict DRA risks

Pooled diagnostic parameters were computed using a random-effects model. This meta-analysis included forest plots of the AUC for machine learning in DRA detection, as well as the sensitivity, specificity, PLR, NLR, DOR, and SROC. The seven included studies' sensitivity ranged from 0.70 to 0.75, while their specificity ranged from 0.84 to 0.92, according to the results of the diagnostic meta-analysis (Fig. 2). These results suggest that machine learning is far more capable of correctly predicting a disease than correctly classifying cases that are not illnesses. Additionally, an assessment was conducted on the machine learning's pooled diagnostic accuracy in predicting DRA. In the pooled analysis, the P value of the Spearman's correlation coefficient was less than 0.05. The sensitivity and specificity I^2 values were 84.25% and 99.15%, respectively, and the chi-square test P-values were all less than 0.05, indicating a significant degree of study heterogeneity. The median outcomes of the pooled predictive data were as follows: 3.62 (95% CI 3.36–3.89) for the pooled PLR, 0.32 (95% CI 0.30–0.35) for the pooled NLR (Fig. 3A, B), and 13.55 (95% CI 11.72–15.67) for the pooled DOR (Fig. 3C). The corresponding SROC curve is displayed in Fig. 3D. The overall SROC curve's AUC value was 0.81, indicating a reasonably high level of accuracy for DRA prediction using machine learning.

Clinical value analysis

The differences in clinical utility between machine learning methods for prediction of DRA were evaluated using Fagan plot analysis. The probability increased from 20 to 62% when the machine learning methods assays were positive and decreased to 7% when the results were negative (Fig. 3E).

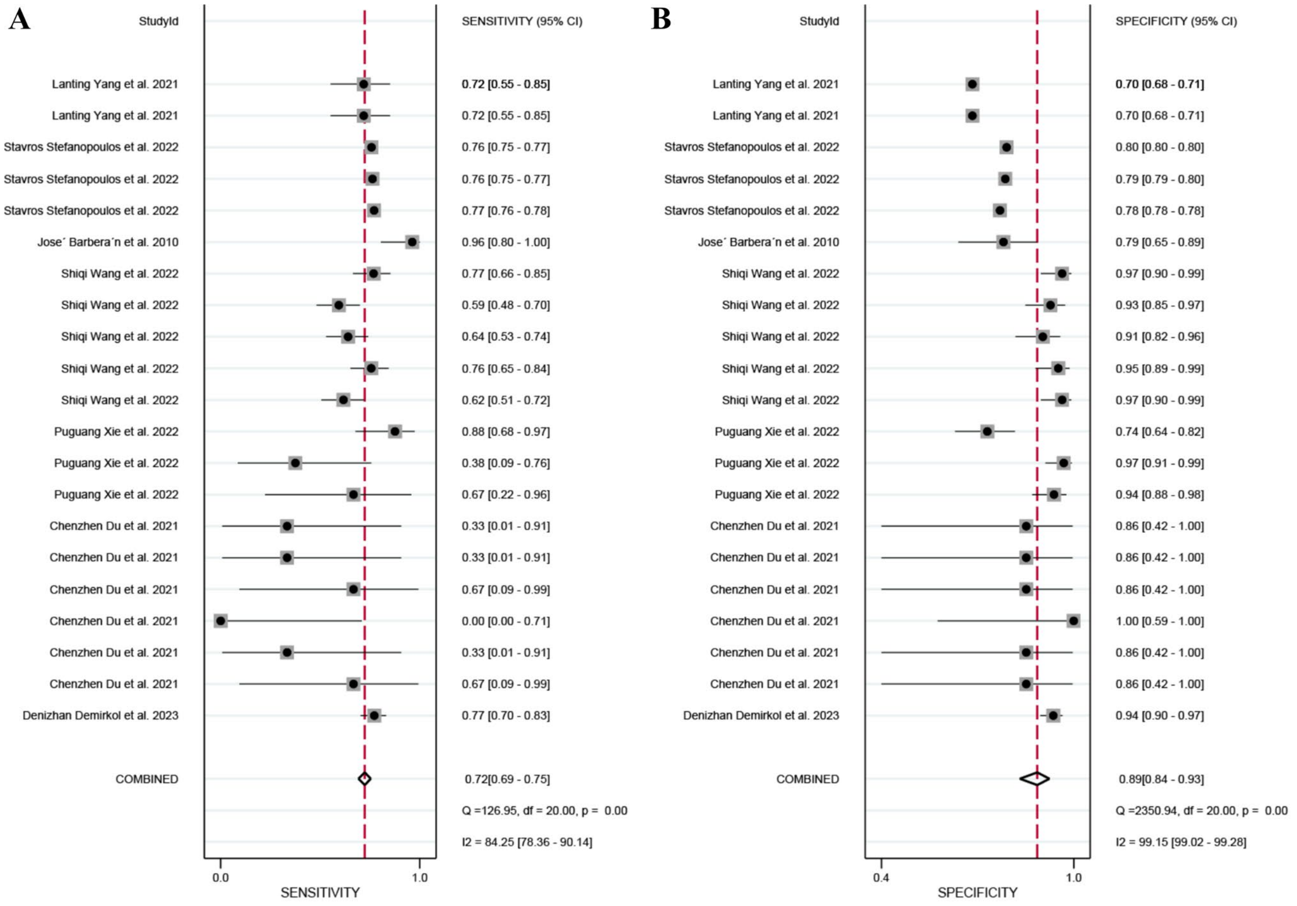

Fig. 2 Machine learning approaches can accurately predict DRA risks.. Forest plots of sensitivity (**A**) and specificity (**B**)

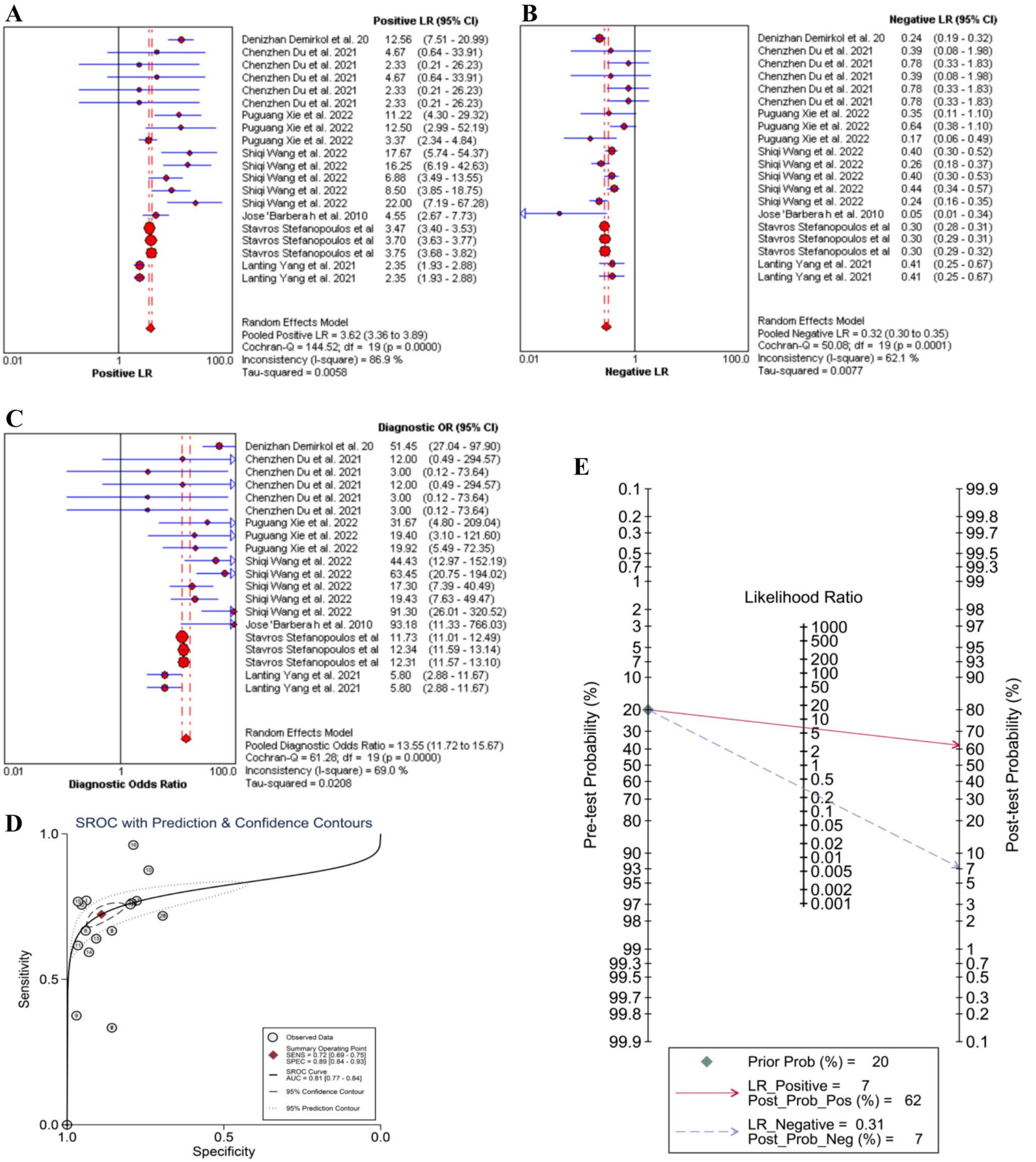

Fig. 3 Comprehensive performance and clinical value of machine learning in the prediction of DRA. Forest plots of PLR (**A**), NLR (**B**), and DOR (**C**). (**D**) SROC curve. **E** Fagan plot

Subgroup analysis and Meta-regression

The heterogeneity of the studies was assessed using bivariate boxplots. As shown in Fig. 4A, four studies (4, 6, 7, 17) [15, 31] were not included in the boxplots. This suggests that these four studies may be the underlying cause of heterogeneity. After carefully reading these four studies, three key influencing factors of heterogeneity were

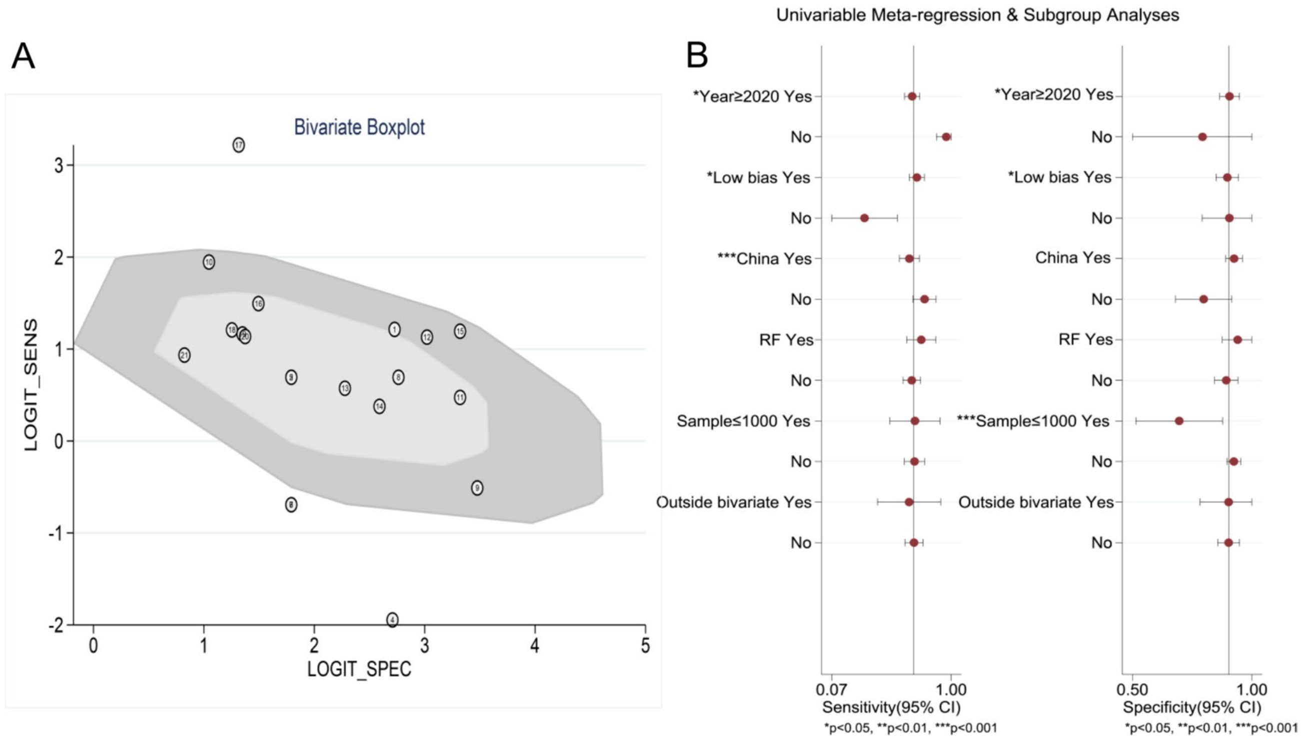

Fig. 4 The source of heterogeneity. **A** bivariate boxplot. **B** univariable meta-regression analysis

identified: machine learning type, sample size, and year of publication. Subsequently, we isolated these four studies from all studies to create separate subgroups for analysis. As shown in Fig. 4B, types of machine learning were not sources of heterogeneity, either in sensitivity or specificity ($P > 0.05$). However, the level of bias and year of publication were all sources of heterogeneity in sensitivity and specificity ($P < 0.05$).

Sensitivity analysis and Publication bias

As shown in Fig. 5A–D, the impact analysis revealed three outlier studies (1, 14, 15) [13, 16], whereas no outlier studies were found in the outlier detection. After removing three studies, the I^2 values for sensitivity and specificity heterogeneity decreased from 67.23 to 87.48% and from 98.65 to 99.06%, respectively (Supplementary Fig. 2). However, the pooled diagnostic accuracy measures were comparable to the overall study (sensitivity: 0.71 vs. 0.72; Specificity: 0.87 vs. 0.89; AUC: 0.83 vs. 0.81), indicating that our results were relatively robust and not significantly influenced by any individual study. In order to assess potential publication bias, we performed the Deeks' funnel plot asymmetry test in our meta-analysis (Fig. 5E). $P = 0.07$ indicated little chance of publication bias among the studies.

GRADE Evidence Quality Assessment

In addition, we applied GRADE assessment method to assess the strength of evidence. Initial level of evidence: All studies had an observational design and the starting grade was "low". Downgrade factors: (1) Risk of bias: two studies had high risk of bias (level -1); (2) Inconsistency: sensitivity and specificity $I^2 > 75\%$ (level -1); (3) Imprecision: the confidence interval of the combined effect size crossed the clinical decision threshold (level -1). Final level of evidence: Very low. These summaries are detailed in Supplementary Table 2.

Discussion

Machine learning has become an inevitable trend in medicine and has already proficient utilization in certain medical domains [32–34]. The study confirmed that machine learning has moderate accuracy in the diagnosis of DRA ($AUC = 0.81$). Furthermore, the combined sensitivity and specificity were 0.72 and 0.89, respectively. In fact, the ideal diagnostic tools should distinguish not only patients with DRA but also other similar diseases, such as traumatic amputation [35, 36]. Machine learning may be effective in ruling out non-DRA (distinguishing traumatic from vascular

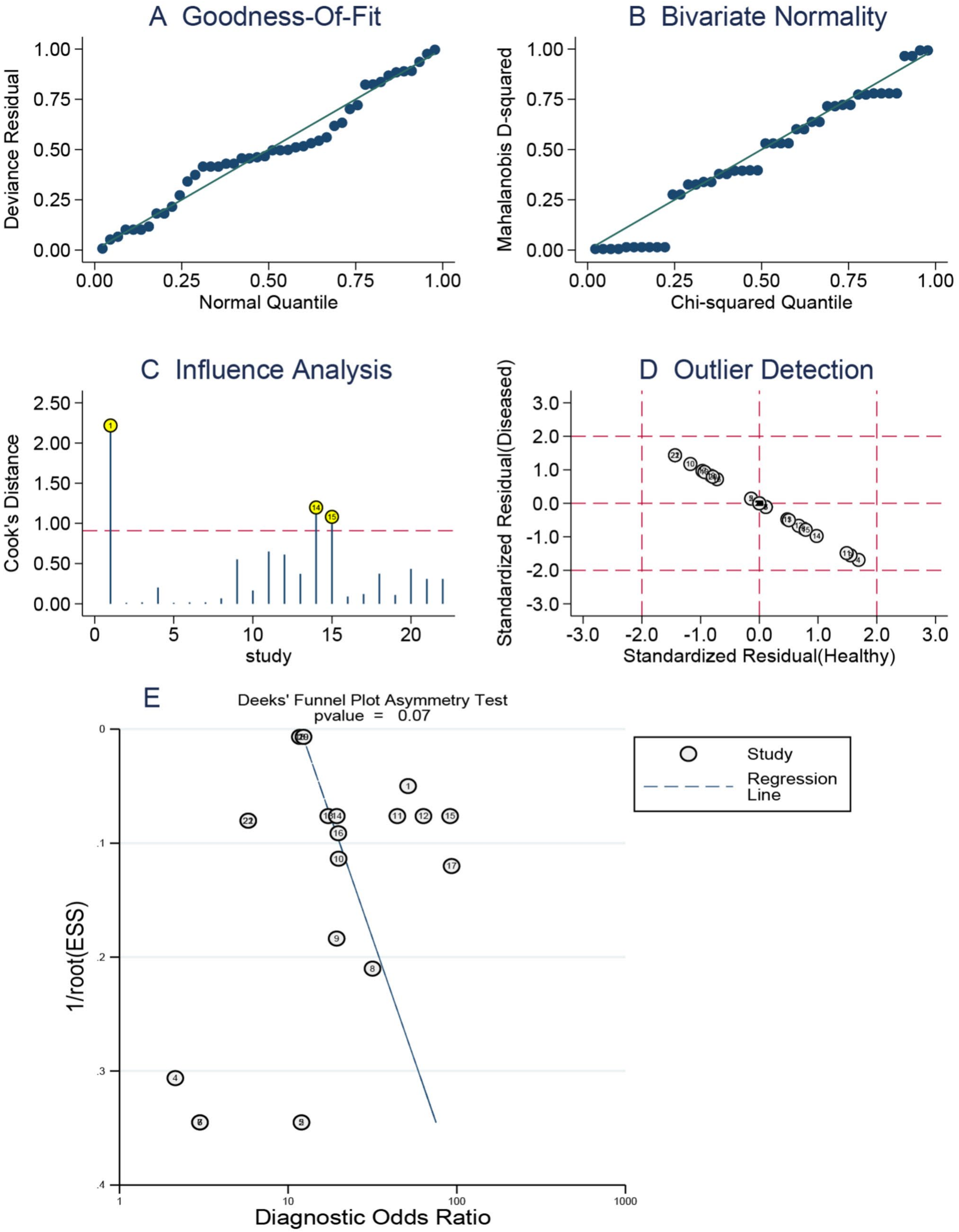

Fig. 5 Sensitivity analysis and bias of publication. **A** Goodness of fit. **B** Bivariate normality. **C** Influence analysis. **D** Outlier detection. **E** The asymmetry test of Deeks' funnel plot for publication

amputations) and is valuable in reducing unnecessary further testing.

However, the sensitivity of 0.72 is slightly lower, suggesting that additional diagnostic methods may be needed in clinical practice. These results are consistent with the trend of previous single-center studies [37], but there are differences in specific performance measures. For example, the amputation prediction system developed by Du et al. based on the XGBoost model [15] reported an accuracy rate of 80%, which is in line with the sensitivity (0.73) and specificity (0.89) shown in this study. However, this accuracy is lower than the 84% reported by Stefanopoulos et al. based on the decision tree model [14]. This difference may be due to the influence of algorithm selection and data heterogeneity: ensemble learning models such as XGBoost have advantages in capturing feature interactions, while single decision trees (DT) are susceptible to sample bias [38, 39], which leads to overestimation of performance in some studies.

This study found that machine learning had significantly higher discrimination ability for non-diabetic amputations than traditional methods, which is consistent with the findings of Alsaade et al. [10] in the diagnosis of skin lesions, showing that machine learning can improve specificity through multi-modal data integration [10]. However, the problem of relatively insufficient sensitivity (0.73) echoes the findings of Wang et al. [16] on Texas grade 3 diabetic foot ulcers [16]. This phenomenon may be due to insufficient feature learning caused by small samples.

The likelihood ratio (LR) is an independent indicator of response authenticity [40]. The PLR of diagnostic tests can reflect their accuracy, and the NLR can reflect the degree of missed diagnosis of diagnostic tests [41–43]. In our meta-analysis, the pooled PLR was 3.62 and NLR was 0.32. Machine learning was 3.62 times more accurate in predicting DRA than the control group, but missed the diagnosis 32 percent of the time. Notably, recent evidence from oncodiagnostics has shown that the PLR value of an optimized machine learning model for preoperative prediction of microvascular invasion in HCC has reached 5.14 [44], highlighting the potential for algorithmic optimization by integrating clinical decision thresholds. This comparison highlights the opportunity to enhance DRA prediction by incorporating domain-specific clinical parameters into model architecture.

As an independent indicator of morbidity, the DOR reflects the degree of correlation between diagnosis and diseases [45]. However, there is still no consensus on how big DOR is. The pooled DOR of this study (DOR = 13.55) was higher than the DOR of serum markers for the diagnosis of any liver fibrosis (AnF) in Sergio et al.'s meta-analysis (DOR = 5.61) [46], but there was still a gap with the high-performance AI model in the diagnosis of malignant tumors (such as tumor bone metastases DOR = 58) [47]. This difference may be due to the multi-factor pathogenesis of DRA. Compared with the single biomarker, the risk of DRA is affected by multi-dimensional indicators such as blood glucose control, infection degree, and vascular status, which increases the difficulty of modeling.

Meta-regression analysis showed that sample size and publication year were the main sources of heterogeneity, which was consistent with the evolution trend of multimodal data integration research. In the field of DRA, early single-center studies (e.g., Barberan et al.'s [31] amputation prediction model based on data from a Spanish single-center) [31] mainly relied on wound imaging features, while recent studies (e.g., Demirkol et al. 2024) [13] have begun to integrate multi-dimensional data such as glycated hemoglobin and inflammatory biomarkers from electronic health records. It is worth noting that heterogeneity may be aggravated by differences in algorithm architecture—such as the XGBoost model developed by Du et al. 2022 [15], while European and American studies have used the decision tree model (CTREE) developed by Stefanopoulos et al. 2022 [14].

This study has several limitations. Firstly, although there is a substantial body of literature on the use of machine learning for predicting DRA, many studies lack essential experimental data, such as TP, TN, and F1 scores. Consequently, only seven articles met the inclusion criteria, which may have compromised the robustness of the findings. Secondly, a comprehensive literature review revealed that most studies were based on small sample sizes and conducted at single centers, potentially limiting the capacity of machine learning to reliably assess DRA. Thirdly, of the seven studies included, three originated from China and two from the USA, with the research being published exclusively in English and Chinese. This may introduce selection bias and diminish the validity of the research outcomes. Lastly, due to the low sensitivity of Deek's funnel plot and the limited number of studies included in this meta-analysis, the possibility of publication bias cannot be entirely ruled out [48].

In conclusion, this study showed that machine learning accurately predicts DRA. Given the limited number of available studies, it is necessary to further examine the validity and potential applicability of machine learning as a predictive indicator of DRA through multi-center, large-sample and prospective studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10238-025-01697-w.

Acknowledgements We thank all authors and reviewers who participated in this work.

Author contributions It was written by Zhigang Chen, Xinliang Liu, and Simeng Li. The manuscript was revised and subjected to critical discussion by Zhenheng Wu, Haifen Tan and Fuqian Yu. Ideas and themes were created by Dongmei Wang and Yawen Bo. All authors have read and consented to publish this manuscript.

Funding This work was supported by grants from National Natural Science Foundation of China (82273232), Changzhou Sci&Tech Program (LCQYBS202309, 2024CZBJ020), Changzhou Medical Center of Nanjing Medical University Program (CZKY102RC202301).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval Not applicable.

Patient consent for publication Not applicable.

References

1. Li Y, Leng Y, Liu Y, et al. Advanced multifunctional hydrogels for diabetic foot ulcer healing: active substances and biological functions. J Diabetes. 2024;16: e13537. https://doi.org/10.1111/1753-0407.13537.
2. Chen C, Li X, Hu Y, et al. Electrical stimulation promoting the angiogenesis in diabetic rat perforator flap through attenuating oxidative stress-mediated inflammation and apoptosis. PeerJ. 2024;12: e16856. https://doi.org/10.7717/peerj.16856.
3. Andjic M, Bozin B, Draginic N, et al. Formulation and evaluation of helichrysum italicum essential oil-based topical formulations for wound healing in diabetic rats. Pharmaceuticals (Basel). 2021. https://doi.org/10.3390/ph14080813.
4. Mineoka Y, Ishii M, Hashimoto Y, et al. Nutritional status assessed with objective data assessment correlates with a high-risk foot in patients with type 2 diabetes. J Clin Med. 2022. https://doi.org/10.3390/jcm11051314.
5. Kerstan A, Dieter K, Niebergall-Roth E, et al. Translational development of abcb5(+) dermal mesenchymal stem cells for therapeutic induction of angiogenesis in non-healing diabetic foot ulcers. Stem Cell Res Ther. 2022;13:455. https://doi.org/10.1186/s13287-022-03156-9.
6. Barjasteh A, Kaushik N, Choi EH, et al. Cold atmospheric pressure plasma: a growing paradigm in diabetic wound healing-mechanism and clinical significance. Int J Mol Sci. 2023. https://doi.org/10.3390/ijms242316657.
7. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, et al. Quality, usability, and effectiveness of mhealth apps and the role of artificial intelligence: current scenario and challenges. J Med Internet Res. 2023;25: e44030. https://doi.org/10.2196/44030.
8. Sun H, Yang H, Mao Y. Personalized treatment for hepatocellular carcinoma in the era of targeted medicine and bioengineering. Front Pharmacol. 2023;14:1150151. https://doi.org/10.3389/fphar.2023.1150151.
9. Gao Y, Wang M, Zhang G, et al. Cluster-based ensemble learning model for aortic dissection screening. Int J Environ Res Public Health. 2022. https://doi.org/10.3390/ijerph19095657.
10. Alsaade FW, Aldhyani T, Al-Adhaileh MH. Developing a recognition system for diagnosing melanoma skin lesions using artificial intelligence algorithms. Comput Math Methods Med. 2021;2021:9998379. https://doi.org/10.1155/2021/9998379.
11. MacMath D, Chen M, Khoury P. Artificial intelligence: exploring the future of innovation in allergy immunology. Curr Allergy Asthma Rep. 2023;23:351–62. https://doi.org/10.1007/s11882-023-01084-z.
12. Xu Y, Hu M, Liu H, et al. A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. Npj Digit Med. 2021;4:48. https://doi.org/10.1038/s41746-021-00417-4.
13. Demirkol D, Erol CS, Tannier X, et al. Prediction of amputation risk of patients with diabetic foot using classification algorithms: a clinical study from a tertiary center. Int Wound J. 2024;21: e14556. https://doi.org/10.1111/iwj.14556.
14. Stefanopoulos S, Qiu Q, Ren G, et al. A machine learning model for prediction of amputation in diabetics. J Diabetes Sci Technol. 2022. https://doi.org/10.1177/19322968221142899.
15. Du C, Li Y, Xie P, et al. The amputation and mortality of inpatients with diabetic foot ulceration in the covid-19 pandemic and postpandemic era: a machine learning study. Int Wound J. 2022;19:1289–97. https://doi.org/10.1111/iwj.13723.
16. Wang S, Wang J, Zhu MX, et al. Machine learning for the prediction of minor amputation in university of texas grade 3 diabetic foot ulcers. PLoS One. 2022;17: e278445. https://doi.org/10.1371/journal.pone.0278445.
17. Xie P, Li Y, Deng B, et al. An explainable machine learning model for predicting in-hospital amputation rate of patients with diabetic foot ulcer. Int Wound J. 2022;19:910–8. https://doi.org/10.1111/iwj.13691.
18. Merzbacher C, Oyarzun DA. Applications of artificial intelligence and machine learning in dynamic pathway engineering. Biochem Soc Trans. 2023;51:1871–9. https://doi.org/10.1042/BST20221542.
19. Casas K, DiPede L, Toema S, et al. Assessing teledentistry versus in-person examinations to detect dental caries: a systematic review and meta-analysis. Jdr Clin Trans Res. 2025. https://doi.org/10.1177/23800844251320974.
20. Yan G, Wang Y, Chen L. Diagnostic performance of artificial intelligence based on biparametric mri for clinically significant prostate cancer: a systematic review and meta-analysis. Acad Radiol. 2025. https://doi.org/10.1016/j.acra.2025.02.044.
21. Singh P, Singhal T, Parida GK, et al. Diagnostic performance of fapi pet/ct vs. (18)f-fdg pet/ct in evaluation of liver tumors: a systematic review and meta-analysis. Mol Imaging Radionucl Ther. 2024;33:77–89. https://doi.org/10.4274/mirt.galenos.2024.99705.
22. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58:882–93. https://doi.org/10.1016/j.jclinepi.2005.01.016.
23. Zhang XY, Li YQ, Yin ZH, et al. Supplements for cognitive ability in patients with mild cognitive impairment or alzheimer's disease: a protocol for systematic review and network meta-analysis of

randomised controlled trials. BMJ Open. 2024;14: e77623. https://doi.org/10.1136/bmjopen-2023-077623.

24. Huang L, Ma L, Zhou Q et al (2024) Accuracy of mri-based radiomics in diagnosis of placenta accreta spectrum: a prisma systematic review and meta-analysis. Med Sci Monit 30:e943461. https://doi.org/10.12659/MSM.943461
25. Shi NN, Li J, Liu GH et al (2024) Artificial intelligence for the detection of glaucoma with sd-oct images: a systematic review and meta-analysis. Int J Ophthalmol 17:408–419. https://doi.org/10.18240/ijo.2024.03.02
26. Shahid S, Iqbal M, Saeed H, et al. Diagnostic accuracy of apple watch electrocardiogram for atrial fibrillation: a systematic review and meta-analysis. Jacc Adv. 2025;4: 101538. https://doi.org/10.1016/j.jacadv.2024.101538.
27. (2024) Robotic-assisted surgery for rectal cancer: an expedited summary of the clinical evidence. Ont Health Technol Assess Ser 24:1–45
28. Zheng X, Li W, Yan Y, et al. Association between the dietary inflammatory index and fracture risk in older adults: a systematic review and meta-analysis. J Int Med Res. 2024;52:645658329. https://doi.org/10.1177/03000605241248039.
29. Hsu CY, Saver JL, Ovbiagele B, et al. Association between magnitude of differential blood pressure reduction and secondary stroke prevention: a meta-analysis and meta-regression. Jama Neurol. 2023;80:506–15. https://doi.org/10.1001/jamaneurol.2023.0218.
30. Liu H, Guo N, Zheng Q, et al. Association of interleukin-6, ferritin, and lactate dehydrogenase with venous thromboembolism in COVID-19: a systematic review and meta-analysis. Bmc Infect Dis. 2024;24:324. https://doi.org/10.1186/s12879-024-09205-3.
31. Barberan J, Granizo JJ, Aguilar L, et al. Predictive model of short-term amputation during hospitalization of patients due to acute diabetic foot infections. Enferm Infecc Microbiol Clin. 2010;28:680–4. https://doi.org/10.1016/j.eimc.2009.12.017.
32. Jeong GH. Artificial intelligence, machine learning, and deep learning in women's health nursing. Korean J Women Health Nurs. 2020;26:5–9. https://doi.org/10.4069/kjwhn.2020.03.11.
33. Malakul W, Seenak P, Jumroon N, et al. Novel coconut vinegar attenuates hepatic and vascular oxidative stress in rats fed a high-cholesterol diet. Front Nutr. 2022;9: 835278. https://doi.org/10.3389/fnut.2022.835278.
34. He Q, Xue Y. Research on the influence of digital finance on the economic efficiency of energy industry in the background of artificial intelligence. Sci Rep. 2023;13:14984. https://doi.org/10.1038/s41598-023-42309-5.
35. Ding L, Cao S, Qu C, et al. Ratiometric crispr/cas12a-triggered cha system coupling with the msre to detect site-specific dna methylation. Acs Sens. 2024;9:1877–85. https://doi.org/10.1021/acssensors.3c02571.
36. Achtnichts L, Gonen O, Rigotti DJ, et al. Global n-acetylaspartate concentration in benign and non-benign multiple sclerosis patients of long disease duration. Eur J Radiol. 2013;82:e848–52. https://doi.org/10.1016/j.ejrad.2013.08.037.
37. Sun S, Chen C, Sheng Z, et al. The distal tibiofibular joint effusion may be a reliable index for diagnosing the distal tibiofibular syndesmosis instability in ankle. Skeletal Radiol. 2024;53:329–38. https://doi.org/10.1007/s00256-023-04395-4.
38. Wan X, Wang Y, Liu Z, et al. Development of an interpretable machine learning model based on ct radiomics for the prediction of post acute pancreatitis diabetes mellitus. Sci Rep. 2025;15:1985. https://doi.org/10.1038/s41598-025-86290-7.
39. Singh Y, Farrelly C, Hathaway QA et al (2024) Persistence landscapes: charting a path to unbiased radiological interpretation. Oncotarget 15:790–792. https://doi.org/10.18632/oncotarget.28671
40. Tang DY, Mao YJ, Zhao J, et al. Seei: spherical evolution with feedback mechanism for identifying epistatic interactions. BMC Genom. 2024;25:462. https://doi.org/10.1186/s12864-024-10373-4.
41. Liu M, Meng K, Jiang J, et al. Comparison of serodiagnosis methods for community-acquired mycoplasma pneumoniae respiratory tract infections in children. Medicine (Baltimore). 2023;102: e34133. https://doi.org/10.1097/MD.0000000000034133.
42. Huang QY, Li PC, Yue JR. Diagnostic performance of serum galactomannan and beta-d-glucan for invasive aspergillosis in suspected patients: a meta-analysis. Medicine (Baltimore). 2024;103: e37067. https://doi.org/10.1097/MD.0000000000037067.
43. Brouwers J, Willems SA, Goncalves LN, et al. Reliability of bedside tests for diagnosing peripheral arterial disease in patients prone to medial arterial calcification: a systematic review. Eclinicalmedicine. 2022;50: 101532. https://doi.org/10.1016/j.eclinm.2022.101532.
44. Zhang J, Huang S, Xu Y, et al. Diagnostic accuracy of artificial intelligence based on imaging data for preoperative prediction of microvascular invasion in hepatocellular carcinoma: a systematic review and meta-analysis. Front Oncol. 2022;12: 763842. https://doi.org/10.3389/fonc.2022.763842.
45. Qian S, Zhang S, Lu M, et al. The accuracy of screening tools for sarcopenia in older chinese adults: a systematic review and meta-analysis. Front Public Health. 2024;12:1310383. https://doi.org/10.3389/fpubh.2024.1310383.
46. Lopez TS, Ayala CO, Ruggiro PB, et al. Accuracy of prognostic serological biomarkers in predicting liver fibrosis severity in people with metabolic dysfunction-associated steatotic liver disease: a meta-analysis of over 40,000 participants. Front Nutr. 2024;11:1284509. https://doi.org/10.3389/fnut.2024.1284509.
47. Tao H, Hui X, Zhang Z, et al. Accuracy of artificial intelligence in detecting tumor bone metastases: a systematic review and meta-analysis. BMC Cancer. 2025;25:286. https://doi.org/10.1186/s12885-025-13631-0.
48. Agasthi P, Kanmanthareddy A, Khalil C, et al. Comparison of computed tomography derived fractional flow reserve to invasive fractional flow reserve in diagnosis of functional coronary stenosis: a meta-analysis. Sci Rep. 2018;8:11535. https://doi.org/10.1038/s41598-018-29910-9.