BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Explainable machine learning for differential diagnosis of diabetic foot infection and osteomyelitis: a two-center study and clinically applicable web calculator using routine blood biomarkers

Parhat Yasin[1†], Shiming Dong[3†], Zubaidanmu Aizezi[2], Yasen Yimit[4], Alimujiang Yusufu[1], Maihemuti Yakufu[1*] and Xinghua Song[1*]

Abstract

Background Diabetic foot complications, including infections and osteomyelitis, pose significant health risks, with high prevalence and amputation rates. Differentiating diabetic foot infection (DFI) from osteomyelitis (OM) is challenging due to overlapping symptoms and limitations of current diagnostic methods. This study aimed to develop and validate an explainable machine learning (ML) model using routine blood biomarkers to improve differential diagnosis and provide a clinically accessible tool.

Methods This retrospective, two-center study included 3,612 patients diagnosed with either DFI ($n=1,699$) or OM ($n=1,913$). Data from Center 1 ($n=3271$) were used for model development (75% training, 25% internal validation), and data from Center 2 ($n=341$) served as an independent external validation cohort. A robust feature selection pipeline identified the most predictive routine biomarkers. Multiple machine learning classifiers were trained and evaluated, with the top-performing model selected based on the area under the receiver operating characteristic curve (AUC), Brier score, and other key metrics. Explainable AI (XAI) techniques (SHAP, LIME) were used to ensure model transparency. A web-based calculator was developed for clinical translation.

Results A LightGBM model using only six biomarkers—Age, HbA1c, Creatinine, Albumin, ESR, and Sodium—was selected as the final model. It achieved an AUC of 0.879 (95% CI 0.854–0.902) in internal validation and demonstrated excellent, generalizable performance in the external cohort with an AUC of 0.942 (95% CI 0.936–0.950). The model was well-calibrated and showed significant clinical utility in decision curve analysis. SHAP analysis quantified the

[†]Parhat Yasin and Shiming Dong contributed equally to this work.

*Correspondence:
Maihemuti Yakufu
mhmtykf@xjmu.edu.cn
Xinghua Song
songxinghua19@163.com

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

specific contribution of each biomarker to individual predictions, enhancing interpretability. The final model was deployed as a user-friendly, publicly accessible web calculator.

Conclusions An externally validated machine learning model based on six routine blood biomarkers can accurately and reliably differentiate DFI from OM. The model demonstrated high discriminative performance and clinical utility. Deployed as a transparent web calculator with integrated explainable AI, this low-cost tool has the potential to aid clinicians in diagnostic decision-making, particularly in resource-limited settings.

Clinical trial number Not applicable.

Keywords Diabetic foot infection, Osteomyelitis, Blood biomarkers, Machine learning, Explainable artificial intelligence (XAI)

Introduction

Diabetic foot complications (DFC) represented a major and escalating challenge in diabetes care, carrying severe consequences for patients. The lifetime risk of developing diabetic foot ulcers (DFU) ranged from 19% to 34%, underscoring their high prevalence [1]. These complications were associated with significant morbidity, frequently leading to infections, lower-extremity amputations, and increased mortality [2, 3]. Notably, approximately 65% of patients experienced ulcer recurrence within five years of initial healing, highlighting the persistent nature of this condition [4]. Diabetic foot infection (DFI) was a common and serious complication in patients with diabetes, associated with substantial health risks and economic costs. As a leading cause of hospitalization and major amputations in this population, DFIs significantly reduced survival rates, with one-year and five-year overall survival rates dropping to 41.7% and 8.3%, respectively, following major amputation1. These infections also increased the risk of sepsis, recurrent hospitalizations, and further amputations, contributing to elevated healthcare expenditures [5]. The financial burden was exacerbated by prolonged antibiotic regimens, extended hospital stays, and surgical procedures [6].

DFIs posed a significant diagnostic challenge, particularly in distinguishing between soft-tissue infections and osteomyelitis (OM), a distinction critical for guiding treatment decisions such as antibiotic duration and surgical intervention [7, 8]. Clinical overlap between DFIs and OM often rendered symptom-based differentiation unreliable, while the multidisciplinary complexity of diagnosis was heightened by OM's association with elevated amputation risks, prolonged antibiotics, and extended hospitalization [9, 10]. Although clinical signs like a positive probe-to-bone test or elevated inflammatory biomarkers suggested OM, definitive diagnosis typically required imaging or bone biopsy [11]. Current non-invasive imaging modalities lacked sufficient specificity, necessitating a combined assessment of clinical, biochemical, and radiographic findings—with bone biopsy remaining the gold standard despite its limited feasibility. However, bone biopsy with culture maintained its status as the gold standard, enabling definitive pathogen identification and guiding targeted treatment despite its invasive nature [12]. This differentiation directly influenced treatment decisions, where misdiagnosis often led to severe consequences—delayed OM treatment increased amputation risks and prolonged hospitalization, while overdiagnosis triggered unnecessary antibiotics or invasive procedures. The development of precise, non-invasive diagnostic tools emerged as an urgent unmet need to optimize DFI management, improve outcomes, and mitigate the burden of diabetic foot complications. Magnetic Resonance Imaging (MRI) emerged as the preferred advanced imaging method due to its superior diagnostic performance, demonstrating sensitivity and specificity of 96.4% and 83.8%, respectively [13]. Additional modalities like positron emission tomography (PET) and single-photon emission computed tomography (SPECT) also showed promise, with PET exhibiting particularly high specificity [14]. The probe-to-bone (PTB) test remained a practical clinical tool for OM diagnosis, offering reliable sensitivity and specificity in high-risk populations [15]. The current diagnostic landscape for DFO and DFI is fragmented, with each method having its own set of limitations. The reliance on invasive procedures like bone biopsy and the variability in the accuracy of non-invasive tests underscore the need for more accessible, accurate, and non-invasive diagnostic tools. Such tools would not only improve diagnostic accuracy but also reduce the burden on patients and healthcare systems by minimizing the need for invasive procedures and prolonged hospital stays.

Routine blood biomarkers offered a promising yet underutilized resource for managing DFI and osteomyelitis OM. Diabetic patients routinely underwent comprehensive blood panels, providing a cost-effective and readily available data source. These panels included markers such as ESR, CRP, and interleukins, which were previously studied for their diagnostic and predictive potential in osteomyelitis among diabetic patients [16, 17]. Subtle, combined patterns within these biomarkers—potentially overlooked by traditional analysis—might reflect distinct pathophysiological states of DFI and OM. For example,

ESR and CRP levels were linked to reinfection risk and infection severity in diabetic foot ulcers [16]. Machine learning (ML) algorithms were particularly suited for analyzing these biomarkers due to their ability to detect complex, non-linear relationships in multidimensional data. By leveraging ML, researchers could uncover hidden patterns and interactions among biomarkers, improving diagnostic accuracy and prognostic assessment for DFI and OM, ultimately enhancing patient outcomes [18]. Previous research on cardiovascular and stroke risk stratification in DFI patients using deep learning faced significant limitations, including small sample sizes and inadequate external validation, which restricted the generalizability of the findings. These methodological shortcomings undermined both the scientific rigor and clinical applicability of the results [19]. The integration of ML in medicine presents a significant challenge due to the "black box" nature of these models, which often leads to hesitancy among clinicians to trust ML predictions without understanding the underlying reasoning. This lack of transparency is a major barrier to the clinical adoption of ML, especially in high-stakes decisions such as DFI and OM management, where model transparency is essential [20].

The primary aim of this study is to develop and externally validate an interpretable machine learning model based on routine blood biomarkers to distinguish diabetic foot infection from osteomyelitis. Key research questions include: (1) Can a model using only routine biomarkers achieve high accuracy in differentiating DFI from OM? (2) How do explainable AI techniques enhance the interpretability of such a model? (3) What is the model's performance in an independent external cohort? The hypothesis is that a machine learning approach incorporating explainable AI can provide a reliable, non-invasive diagnostic tool that outperforms traditional methods in accuracy and clinical utility.

Methods

This retrospective study received ethical approval from the Institutional Ethics Committees of the the Sixth Affiliated Hospital of Xinjiang Medical University and the First People's Hospital of Kashi Prefecture, and the committees waived the requirement for individual patient informed consent due to the study's retrospective design and the complete anonymization of patient data before analysis. All procedures were conducted in accordance with the Declaration of Helsinki and relevant institutional guidelines.

Cohort definition and data acquisition

We retrospectively identified enrolled patients admitted to the Sixth Affiliated Hospital (designated Center 1) and the First People's Hospital of Kashi Prefecture (designated Center 2) between January 2015 and December 2024. Center 1 data were randomly partitioned, allocating 75% for model training and 25% for internal validation. Data from Center 2 constituted an independent external validation cohort to assess model generalizability. Patient selection adhered to specific criteria outlined in the study flowchart (Fig. 1). This retrospective study included patients who met the following criteria: (1) Aged 18 years or older; (2) Possessed a confirmed diagnosis of diabetes mellitus; (3) Presented at one of the two participating clinical centers with diabetic foot complications necessitating diagnostic evaluation for infection; (4) Had comprehensive routine blood biomarker results; (5) Received a definitive final diagnosis distinguishing between soft tissue DFI alone and OM, based on a composite reference standard integrating clinical findings (e.g., probe-to-bone test), laboratory results, imaging data (primarily MRI when available), microbiological or histopathological evidence from bone biopsy or surgical debridement if performed, and documented clinical course and response to treatment [21, 22]. Conversely, patients were excluded based on the following criteria: (1) Age under 18 years; (2) Absence of a confirmed diabetes mellitus diagnosis; (3) Primary foot pathology confirmed to be unrelated to DFI/OM, such as isolated Charcot neuroarthropathy without superimposed infection suspicion, acute major trauma, or crystalline arthropathy; (4) Presence of concurrent systemic inflammatory or infectious conditions (e.g., active autoimmune disease flare, sepsis originating elsewhere, recent major surgery, active malignancy undergoing treatment) known to significantly confound the interpretation of routine blood biomarkers in the context of foot infection; (5) Incompleteness of the core routine blood biomarker dataset, defined as more than 30% missing values for the variables under investigation during the relevant diagnostic timeframe; (6) Insufficient or ambiguous clinical, imaging, laboratory, or follow-up information preventing a reliable classification into the DFI or OM categories according to the established reference standard.

Biomarker selection and data processing

We collected demographic information (age, gender and a comprehensive panel of routine blood biomarkers typically obtained within the first 24–48 h of hospital admission or initial workup for DFC. These candidate features represented various physiological domains potentially relevant to infection and inflammation in diabetes, including: inflammatory markers (e.g., white blood cell count [WBC] and differential, erythrocyte sedimentation rate [ESR], C-reactive protein [CRP]); glycemic status (glycated hemoglobin [HbA1c]); renal function (e.g., creatinine, estimated glomerular filtration rate [eGFR], uric acid, cystatin C [CYS_C]); hepatic function (e.g., albumin

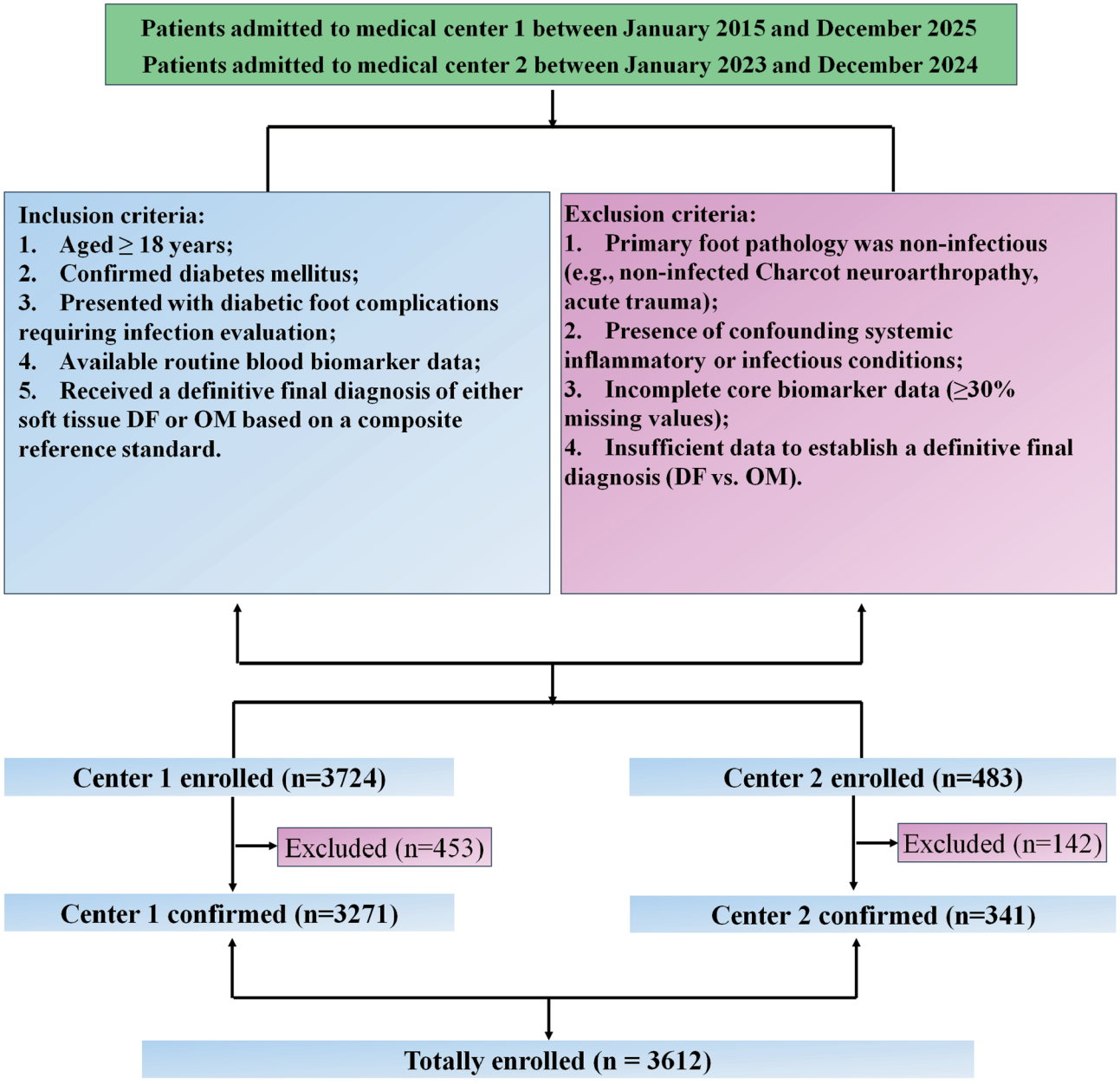

Fig. 1 Inclusion and exclusion process workflow

[ALB], alanine aminotransferase [ALT], aspartate aminotransferase [AST], alkaline phosphatase [ALP]); electrolytes; muscle enzymes (creatine kinase [CK]); and lipid profiles (e.g., triglycerides [TG], high-density lipoprotein [HDL], low-density lipoprotein [LDL]). The full list of evaluated biomarkers is detailed in Supplementary Fig. 1. Prior to analysis, data underwent rigorous preprocessing. Missing biomarker values were addressed using Random Forest imputation, a robust technique that estimates missing data based on observed values across other features [23]. To counteract potential bias introduced by differing scales among biomarkers, all features were standardized using *StandardScaler*. Given the potential imbalance in the prevalence of DFI versus OM cases within clinical cohorts, we applied the Synthetic Minority Oversampling Technique (SMOTE) to the training dataset. This technique synthetically generates new instances of the minority class (OM in this context) based on feature space similarities, creating a balanced dataset for model training and preventing bias towards the majority class. (See Fig. 2)

Feature optimization and machine learning model development

To identify the most parsimonious and informative set of biomarkers for differentiating DFI from OM, we

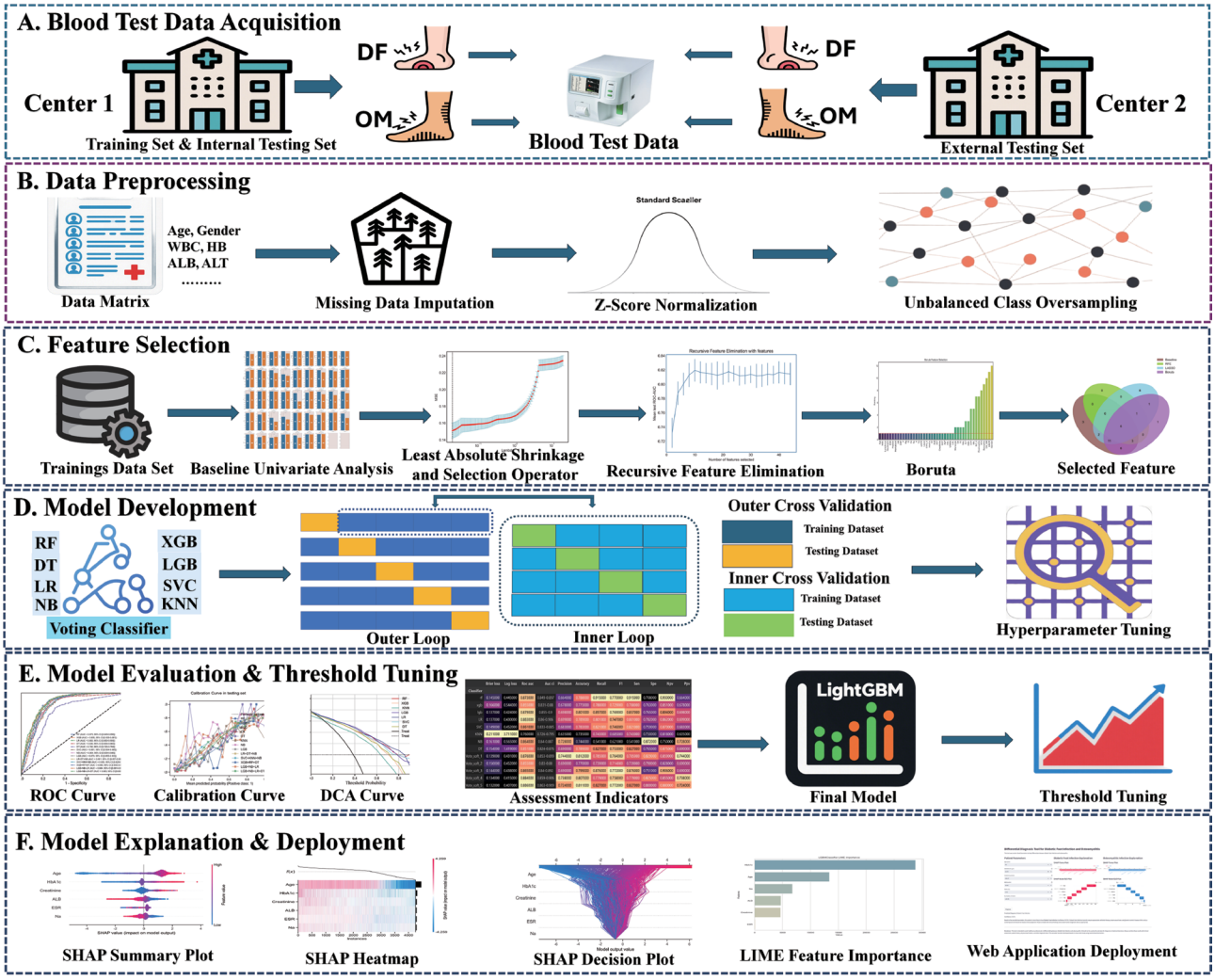

Fig. 2 Study workflow

employed a multi-pronged feature selection strategy on the training data. Three distinct algorithms were utilized: LASSO (Least Absolute Shrinkage and Selection Operator), which performs L1 regularization penalizing feature coefficients towards zero; the Boruta algorithm, which compares feature importance against randomized shadow features; and Recursive Feature Elimination (RFE), which iteratively removes the least contributory features. Only biomarkers consistently identified by the intersection of these three methods were retained for final model construction, ensuring robustness in feature selection. Inter-feature correlations among the selected biomarkers were verified using *Spearman* correlation analysis. Using the selected biomarker set, we developed and compared several supervised machine learning classifiers. These included: logistic regression (LR) as a standard benchmark; advanced gradient boosting machines (extreme gradient boosting [XGBoost] and light gradient boosting machine [LightGBM]); ensemble methods like random forest (RF); a single decision tree (DT); support vector machine classifier (SVC); k-nearest neighbors (KNN); and Gaussian naive Bayes (GNB). Furthermore, we constructed a soft voting ensemble classifier, combining the probabilistic predictions from the top-performing individual models (LR, XGBoost, LightGBM, RF) to potentially enhance predictive accuracy and stability. Model hyperparameters were meticulously optimized using a nested cross-validation approach (5-fold outer loop, 4-fold inner loop) coupled with grid search within the training dataset, aiming to maximize performance while minimizing the risk of overfitting. The hyperparameters for tuning were available in Supplementary File Table 1.

Performance evaluation and interpretability analysis

Model performance was rigorously assessed on both the held-out internal test set (Center 1) and the independent external validation cohort (Center 2). Key discrimination

metrics included the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, F1-score, accuracy, and balanced accuracy. Calibration and probabilistic performance were evaluated using the Brier score and log loss. To identify an optimal probability threshold for clinical application balancing sensitivity and specificity trade-offs, we employed the *TunedThresholdClassifierCV* method (5-fold cross-validation, 50 repeats), specifically optimizing for the F1-score. Recognizing the critical need for transparency in clinical AI tools, particularly for high-stakes decisions like OM diagnosis, we incorporated explainable AI (XAI) techniques. SHapley Additive exPlanations (SHAP) values were calculated to quantify the marginal contribution of each selected biomarker to the model's prediction for both individual patients (local explanation) and the cohort overall (global feature importance) [24]. As a complementary method, Local Interpretable Model-agnostic Explanations (LIME) was used to provide alternative local explanations by approximating the complex model's behavior near specific instances with simpler, interpretable models [25]. These analyses identified the key biomarkers driving the model's differentiation between DFI and OM.

Development of a clinical decision support tool

To translate our findings into a practical clinical tool, we developed a prototype interactive web-based calculator using the Streamlit's Python library [26]. This application allows clinicians to input a patient's values for the selected biomarkers and receive the predicted probability of OM. The interface was designed for ease of use and aimed to potentially incorporate visualizations derived from SHAP values to aid in understanding the prediction rationale, thereby facilitating informed clinical decision-making.

Statistical analysis

All statistical analyses, including normality testing (Shapiro-Wilk test) and descriptive statistics (mean ± SD or median [IQR]), were performed using R version 4.2.1. Machine learning model development, feature selection, evaluation, and XAI analyses were conducted in Python version 3.9.5, utilizing core libraries such as Scikit-learn, XGBoost, LightGBM, SHAP, LIME, and Streamlit. Statistical significance was typically set at $P < 0.05$ where applicable.

Results

Study population

The study cohort comprised 3,612 patients from two medical centers, delineated into two distinct groups for differential diagnosis: 1,699 patients with diabetic foot infection (DF) and 1,913 with osteomyelitis (Osteo). A comprehensive analysis revealed significant demographic and clinical differences between the groups, with DF patients being substantially older (mean age 63.2 vs. 41.7 years, $p < 0.001$) and the group containing a higher proportion of males (71.3% vs. 66.6%, $p = 0.003$). Widespread disparities were also identified across numerous laboratory parameters. Compared to the osteomyelitis group, patients with DF presented with a more pronounced systemic inflammatory profile, characterized by significantly higher levels of white blood cells, neutrophils, C-reactive protein, and erythrocyte sedimentation rate, alongside lower lymphocyte counts (all $P < 0.001$). This was accompanied by evidence of poorer metabolic control and renal function in the DF cohort, underscored by significantly higher HbA1c, creatinine, cystatin C, and uric acid, and a lower estimated glomerular filtration rate (all $p < 0.001$). Significant variations were also noted in electrolytes, liver function markers such as GGT and AST, and lipid profiles. Conversely, several biomarkers, including creatine kinase, alanine aminotransferase, and low-density lipoprotein, did not show statistically significant differences. These extensive baseline disparities underscored the complex, multifactorial nature of these conditions and provided a strong rationale for developing a machine learning model capable of integrating these diverse signals for accurate differential diagnosis. Detailed information on all baseline characteristics and the patient selection process is available in Table 1 and Supplementary Fig. 1, respectively.

Feature engineering

We employed various feature selection techniques to identify the most relevant predictors, emphasizing the need for robust methods that balanced predictive accuracy with model transparency. Figure 3A presented the cross-validation curve for LASSO, which plotted the mean squared error (MSE) against the regularization parameter lambda, helping to identify the optimal lambda that minimized prediction error. Figure 3B illustrated the LASSO coefficient paths, showing how feature coefficients shrank toward zero as the regularization strength (log-alpha) increased, thereby performing feature selection. Figure 3C displayed the Boruta feature selection results, comparing the importance scores of selected features with those of shadow features to identify significant predictors. Figure 3D showed the Recursive Feature Elimination (RFE) process, revealing how model performance changed as features were sequentially removed. Figure 3E displayed a Venn diagram that compared the overlap of features selected by different methods, including Baseline Analysis, LASSO, RFE, and Boruta, emphasizing both shared and unique predictors across strategies. The common predictors identified were ALB, Age, Creatinine, ESR, HbA1c, and Na.

Table 1 Patients baseline characteristics

Characteristics	ALL (*N* = 3612)	DF (*N* = 1699)	Osteo (*N* = 1913)	*P*
Age	51.8 ± 18.9	63.2 ± 11.8	41.7 ± 18.2	<0.001
Gender:				0.003
Female	1126 (31.2%)	488 (28.7%)	638 (33.4%)	
Male	2486 (68.8%)	1211 (71.3%)	1275 (66.6%)	
WBC	9.01 ± 4.65	9.97 ± 5.24	8.15 ± 3.85	<0.001
Lymphocyt	1.91 ± 1.03	1.65 ± 0.70	2.14 ± 1.20	<0.001
Neutrophil	6.23 ± 4.48	7.40 ± 5.17	5.19 ± 3.45	<0.001
Monocyte	0.65 ± 0.32	0.69 ± 0.31	0.61 ± 0.32	<0.001
Basophils	0.03 ± 0.02	0.04 ± 0.02	0.03 ± 0.02	<0.001
Eosinophils	0.16 ± 0.16	0.15 ± 0.18	0.16 ± 0.14	0.004
HB	119 ± 23.3	115 ± 23.5	123 ± 22.5	<0.001
Hematocrit	34.7 ± 10.3	33.8 ± 9.49	35.6 ± 10.9	<0.001
Platelet	306 ± 125	292 ± 116	319 ± 132	<0.001
ESR	37.7 ± 13.8	38.5 ± 10.3	36.9 ± 16.2	<0.001
CRP	38.3 ± 35.7	42.0 ± 33.0	35.0 ± 37.6	<0.001
CK	141 ± 770	158 ± 988	125 ± 502	0.223
CK_MB	16.6 ± 8.67	16.4 ± 4.32	16.8 ± 11.2	0.107
TC	3.76 ± 0.91	3.76 ± 0.97	3.77 ± 0.86	0.727
TG	1.47 ± 0.97	1.57 ± 0.94	1.38 ± 0.99	<0.001
K	4.03 ± 2.03	4.16 ± 2.92	3.92 ± 0.46	0.001
Na	138 ± 6.15	136 ± 7.86	139 ± 3.60	<0.001
Ca	1.94 ± 0.54	1.75 ± 0.60	2.11 ± 0.41	<0.001
Cl	103 ± 5.60	103 ± 7.12	104 ± 3.66	<0.001
Mg	0.86 ± 0.17	0.86 ± 0.23	0.85 ± 0.09	0.106
P	1.27 ± 0.54	1.23 ± 0.71	1.31 ± 0.31	<0.001
Creatinine	107 ± 267	151 ± 371	67.5 ± 99.0	<0.001
eGFR	101 ± 19.9	96.3 ± 17.8	106 ± 20.7	<0.001
Uacid	303 ± 193	329 ± 252	280 ± 114	<0.001
CYS_C	1.07 ± 0.55	1.21 ± 0.68	0.94 ± 0.36	<0.001
HbA1c	2.12 ± 0.40	2.26 ± 0.44	1.99 ± 0.32	<0.001
TP	69.4 ± 7.93	68.7 ± 8.01	70.1 ± 7.80	<0.001
ALB	36.2 ± 6.23	34.0 ± 6.02	38.3 ± 5.69	<0.001
AST	33.6 ± 206	41.5 ± 299	26.5 ± 29.0	0.039
ALT	30.2 ± 142	34.3 ± 201	26.6 ± 47.6	0.126
AFU	20.0 ± 5.09	19.9 ± 3.45	20.2 ± 6.19	0.065
NT_5	6.22 ± 4.32	6.60 ± 4.94	5.89 ± 3.65	<0.001
HDL	0.92 ± 0.29	0.85 ± 0.25	0.98 ± 0.31	<0.001
LDL	2.44 ± 0.73	2.43 ± 0.77	2.45 ± 0.69	0.366
GGT	43.6 ± 58.5	51.6 ± 70.7	36.5 ± 43.9	<0.001
ALP	116 ± 67.3	115 ± 70.9	117 ± 63.9	0.304
LDH	214 ± 393	208 ± 139	219 ± 524	0.393
DBIL	2.32 ± 8.22	1.99 ± 9.84	2.62 ± 6.44	0.023
IBIL	6.70 ± 5.37	6.68 ± 5.00	6.72 ± 5.69	0.837
TBIL	12.1 ± 16.5	13.5 ± 20.5	10.9 ± 11.7	<0.001
UMAlb	363 ± 159	363 ± 206	363 ± 98.8	0.962
UACR	675 ± 111	675 ± 161	675 ± 23.1	0.967

The investigation of feature importance played a crucial role in medical data analysis. Thus, we carried out permutation importance analysis based on the aforementioned selected features among Tree-based algorithms. Figure 4A displayed the permutation importance scores for the Random Forest model, where Age was identified as the most influential feature, followed by ALB and HbA1c. Figure 4B revealed the results for the XGBoost model, with Age again ranking highest, while Creatinine and HbA1c emerged as secondary predictors. Figure 4C illustrated the feature rankings for the LightGBM model, highlighting Age, Na, and HbA1c as key contributors. Figure 4D showed the Decision Tree model's permutation importance, where Age, Na, and HbA1c remained prominent but varied in their relative significance. Across all models, Age consistently appeared as the most critical feature, with HbA1c and Creatinine frequently identified as significant secondary predictors.

Performance of each model

To identify the optimal algorithm, Fig. 5A and B compared model performance between validation and testing datasets using ROC curves. In the training set (Fig. 5A), ensemble models SVC + KNN + NB (AUC = 0.986, 95% CI: 0.983–0.989), XGB + RF + DT (AUC = 0.972, 95% CI: 0.968–0.977), and LGB + NB + LR + DT (AUC = 0.912, 95% CI: 0.904–0.922) achieved the highest AUCs. In the testing set (Fig. 5B), the Light Gradient Boosting (LGB) model demonstrated robust performance with an AUC of 0.879 (95% CI: 0.854–0.902), the highest among individual models, alongside the lowest Brier loss (0.137) and log loss (0.424). The LGB model also exhibited balanced precision (0.698), accuracy (0.801), and recall (0.857). Performance declines in simpler models (e.g., DT: testing AUC = 0.760) highlighted overfitting risks. Based on these metrics, the LGB model was selected as the final algorithm for its consistent discriminative capacity and generalizability. Figure 5C presented the decision curve, showing the net benefit of using the model across different threshold probabilities compared to strategies of treating all or none. The curve demonstrates that the model provided a positive net benefit over a wide range of thresholds, indicating its potential value in clinical decision-making. Figure 5D showed the calibration plot, comparing the model's predicted probabilities against observed outcomes. The plot showed good alignment with the line of perfect calibration, suggesting reliable probability estimates. These visuals collectively highlight the model's clinical utility and predictive reliability, reinforcing its applicability in real-world settings. The detailed evaluation indicators were available in Table 2.

Threshold optimalization

The best-performing LGB model initially demonstrated a relatively lower AUC compared to other models. Figure 6A compared training and testing metrics across multiple indicators, including accuracy, balanced accuracy, F1 score, precision, recall, sensitivity, specificity,

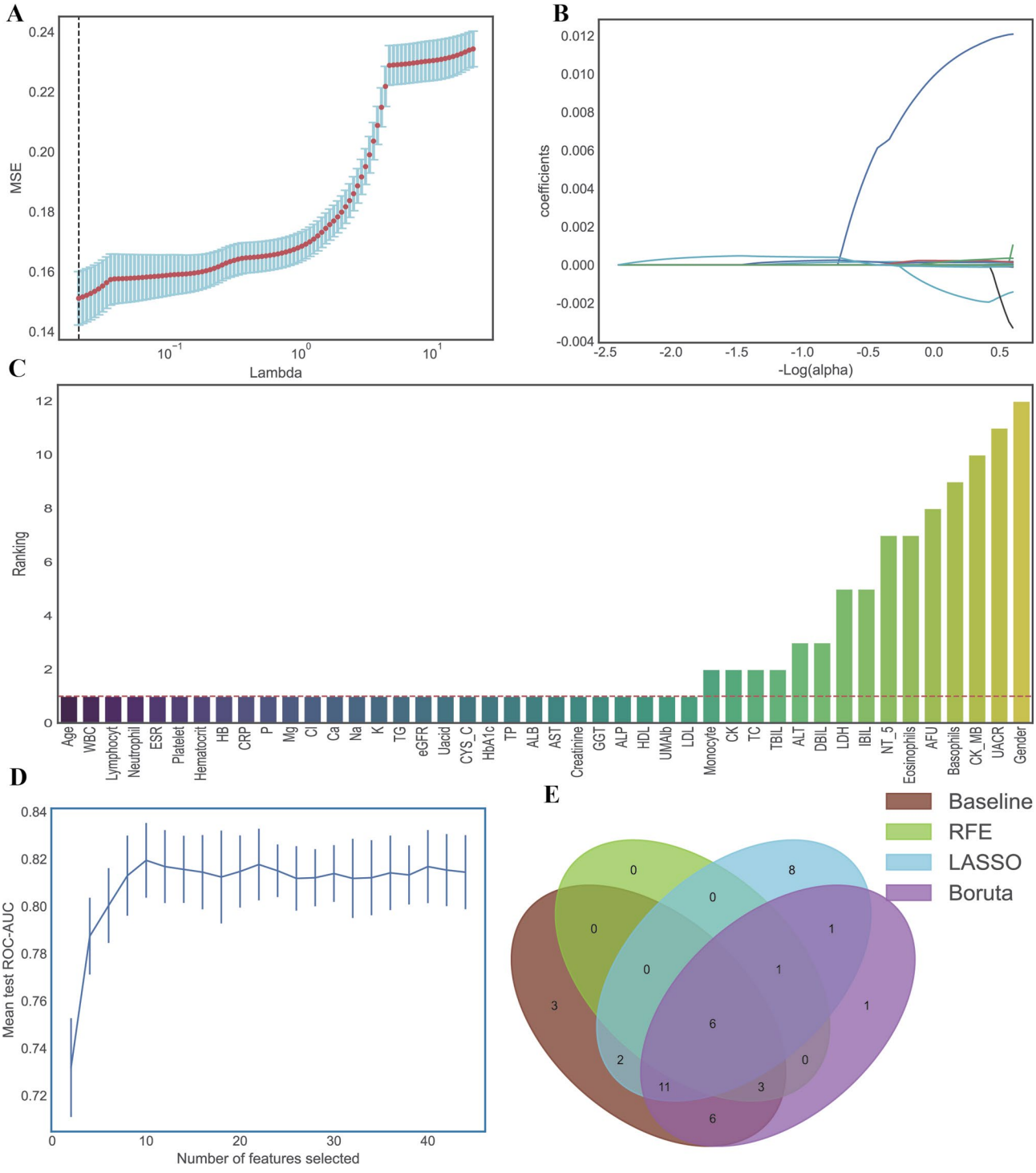

Fig. 3 Variable selection. (**A**): Cross-validation curve for LASSO, plotting MSE against lambda; (**B**): LASSO coefficient paths showing feature shrinkage with increasing regularization; (**C**): Boruta feature selection results, comparing importance scores of selected features; (**D**): Recursive Feature Elimination (RFE) process, showing performance changes with feature removal; (**E**): Venn diagram comparing feature overlap across baseline analysis, LASSO, RFE, and Boruta

NPV, and PPV. The results revealed that training metrics consistently exceeded testing metrics, all indicators were higher than 0.70 in both training and testing datasets. Figure 6B presented a comparison of Brier scores and negative log-loss metrics between training and testing datasets during threshold tuning, with lines connecting corresponding metrics across folds. Figure 6C displayed the distribution of decision thresholds across different

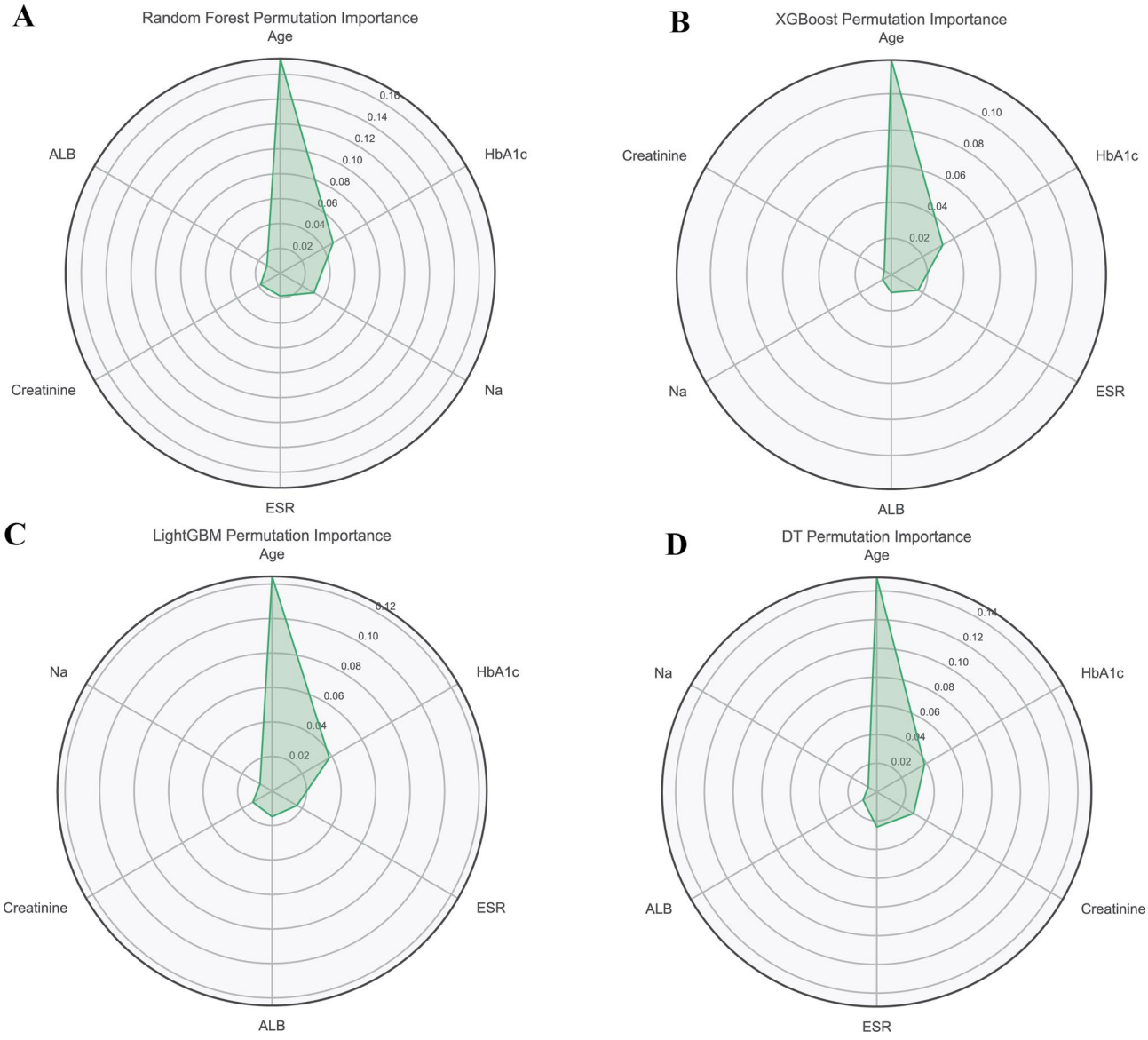

Fig. 4 Permutation importance across models. (**A**): Random Forest permutation importance scores for clinical features; (**B**): XGBoost permutation importance highlighting key predictors; (**C**): LightGBM feature rankings based on permutation importance; (**D**): Decision Tree permutation importance showing feature contributions

cross-validation folds, showing a peak at 0.44, which suggested optimal classification performance at this threshold. These findings addressed the initial concerns about model robustness and emphasized the importance of refining model parameters to ensure clinical applicability.

The final LGB model's performance was evaluated on an external dataset to confirm its generalization ability and clinical applicability. The model exhibited excellent discriminatory power, as demonstrated by the receiver operating characteristic (ROC) curve (Fig. 7A), achieving an AUC of 0.942 (95% CI: 0.936–0.950). The ROC curve was positioned high in the upper-left corner, well above the diagonal line of no-discrimination, indicating a strong ability to differentiate between positive and negative cases. The reliability of the model was assessed using a calibration plot, which revealed strong agreement between predicted probabilities and actual observed frequencies (Fig. 7B). The calibration curve (blue line) closely followed the ideal diagonal line, suggesting that the model's probability outputs were well-calibrated and trustworthy. Additionally, the clinical utility of the model was validated through decision curve analysis (DCA) (Fig. 7C). The DCA showed that using the LGB model to guide clinical decisions provided a significant and positive net benefit across a wide range of threshold probabilities (approximately 0.05 to 0.95), consistently surpassing the alternative strategies of treating all patients (black

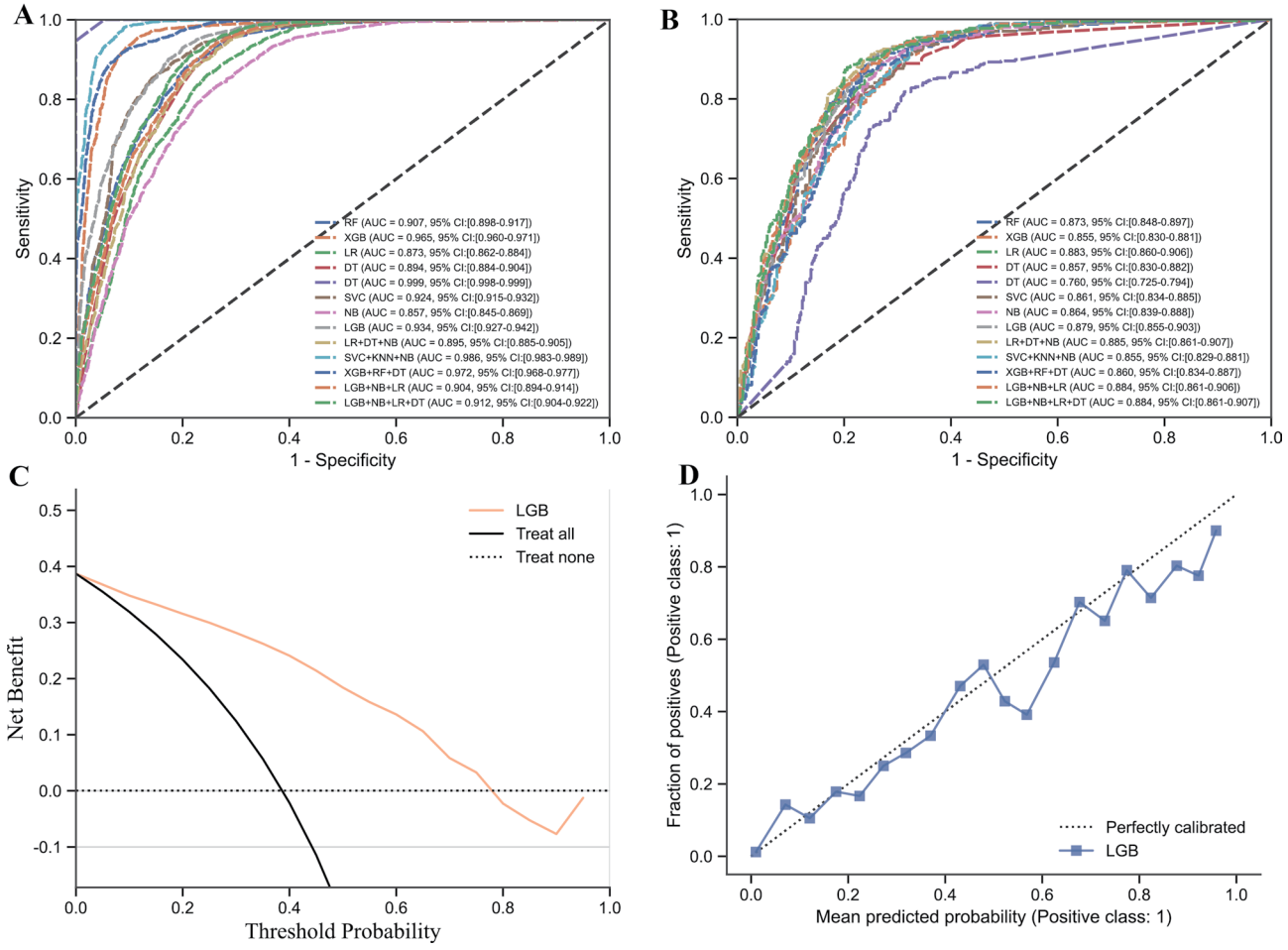

Fig. 5 Comparison of model performance on training and internal testing sets. (**A**): ROC curves for various machine learning models on the training set, including Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), Decision Tree (DT), Support Vector Classifier (SVC), Naive Bayes (NB), LightGBM (LGB), and ensemble models such as LR + DT + NB and SVC + KNN + NB. The curves illustrate the trade-off between sensitivity and specificity, with the area under the curve (AUC) values indicating the overall performance of each model. (**B**): ROC curves for the same set of models on the internal testing set, showing a decrease in performance compared to the training set. The AUC values highlight the variability in model performance between the training and testing sets. (**C**): Decision curve showing net benefit across threshold probabilities in testing dataset; (**D**): Calibration plot comparing predicted and observed probabilities in testing dataset

Table 2 Detailed assessment of each model

Classifier	Brier loss	Log loss	Precision	Accuracy	Recall	F1	Sen.	Spe.	Npv.	Ppv.
RF	0.166	0.544	0.678	0.773	0.788	0.729	0.788	0.763	0.851	0.678
XGB	0.137	0.424	0.698	0.801	0.857	0.769	0.857	0.765	0.894	0.698
LGB	0.137	0.43	0.699	0.789	0.801	0.747	0.801	0.782	0.862	0.699
LR	0.148	0.452	0.683	0.783	0.821	0.746	0.821	0.759	0.87	0.683
SVC	0.211	3.711	0.635	0.735	0.743	0.685	0.743	0.73	0.818	0.635
KNN	0.161	0.565	0.728	0.744	0.541	0.621	0.541	0.872	0.75	0.728
NB	0.155	1.248	0.679	0.779	0.814	0.741	0.814	0.757	0.866	0.679
DT	0.133	0.414	0.748	0.817	0.795	0.771	0.795	0.831	0.865	0.748
Voting mode1 1	0.15	0.452	0.692	0.772	0.739	0.715	0.739	0.792	0.828	0.692
Voting mode1 2	0.146	0.446	0.694	0.801	0.87	0.772	0.87	0.757	0.902	0.694
Voting mode1 3	0.134	0.415	0.738	0.807	0.779	0.758	0.779	0.825	0.855	0.738
Voting mode1 4	0.133	0.414	0.732	0.811	0.808	0.768	0.808	0.813	0.87	0.732
Voting mode1 5	0.166	0.544	0.678	0.773	0.788	0.729	0.788	0.763	0.851	0.678

Model 1: LR + LR + NB, Model 2: SVC + KNN + NB, Model 3: XGB + RF + DT, Model 4: LGB + NB + LR, Model 5: LGB + NB + LR + DT. Sen.: Sensitivity; Spe.: Specificity; NPV: Negative Predictive Value; PPV: Positive Predictive Value

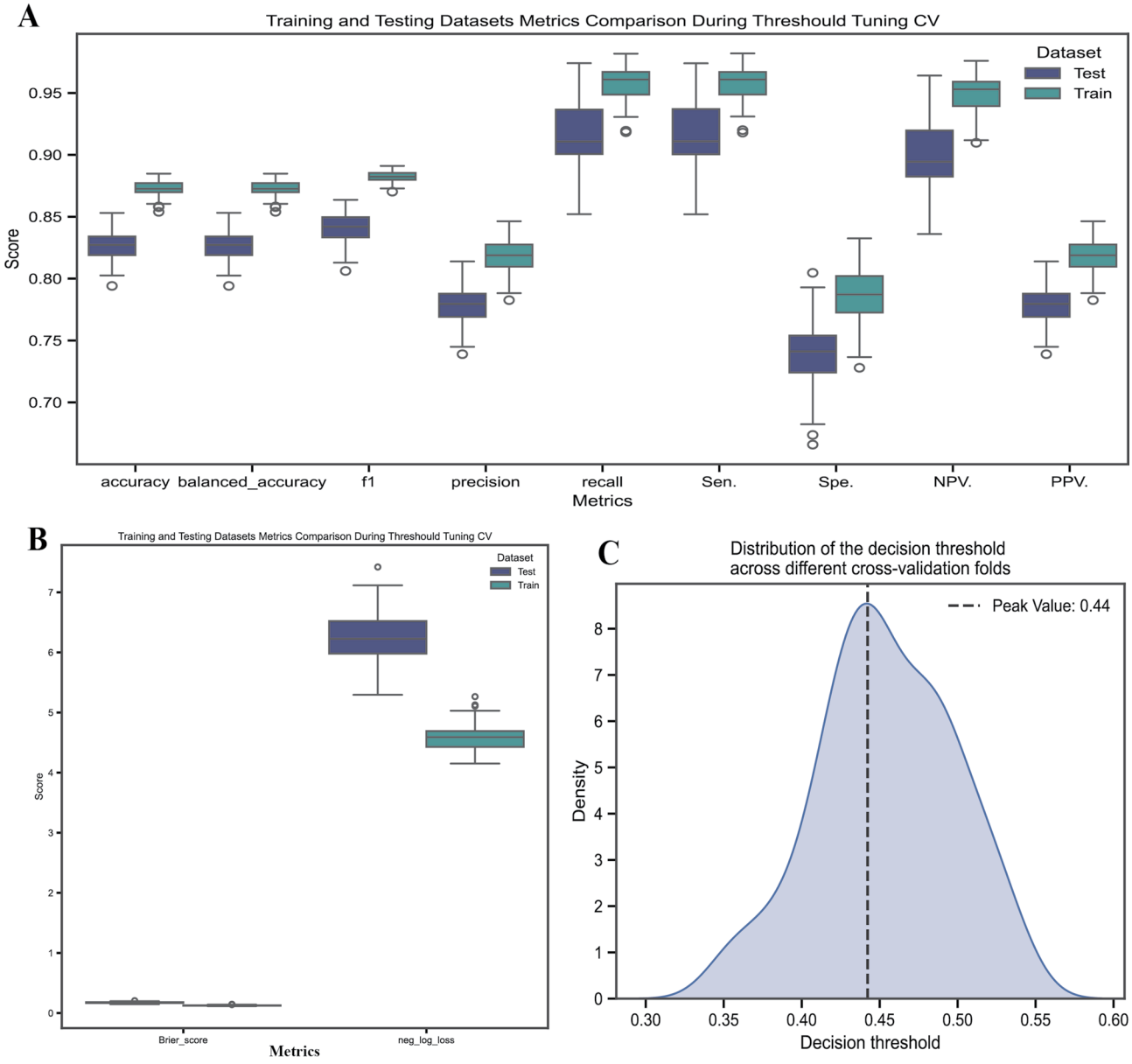

Fig. 6 Comparison of performance metrics between training and testing datasets. (**A**): Box plots comparing various performance metrics between the training and testing datasets, including accuracy, balanced accuracy, F1 score, precision, recall, sensitivity, specificity, NPV, and PPV. (**B**): Box plots showing the Brier score and negative log loss for both the training and testing datasets. (**C**): Density plot of the decision threshold, with a peak value indicating the optimal threshold for classifying new instances

line) or treating none (horizontal grey line). For a detailed comparison with previous work, see Table 3.

Figure 8 demonstrated the stability and robustness of the final model through rigorous 10-fold cross-validation on the overall dataset. The ROC curves for each fold showed consistently high performance with minimal variability, supported by tightly clustered AUC values ranging from 0.88 to 0.91. Specifically, two folds achieved 0.88 (fold 0, 3), two reached 0.89 (fold 4, 7), one attained 0.90 (fold 6), and five achieved the highest AUC of 0.91 (fold 1, 2, 5, 8, 9), reflecting the model's stability. The mean ROC curve produced an overall AUC of 0.90 ± 0.01, with all performance curves positioned far above the baseline chance level (dashed line), confirming the model's strong discriminative power.

Model interpretation

To elucidate the decision-making process of the predictive model, we employed both SHAP and LIME methodologies. The SHAP summary plot revealed the global impact of each feature on the model's output, ranking features by their mean absolute SHAP value (Fig. 9A).

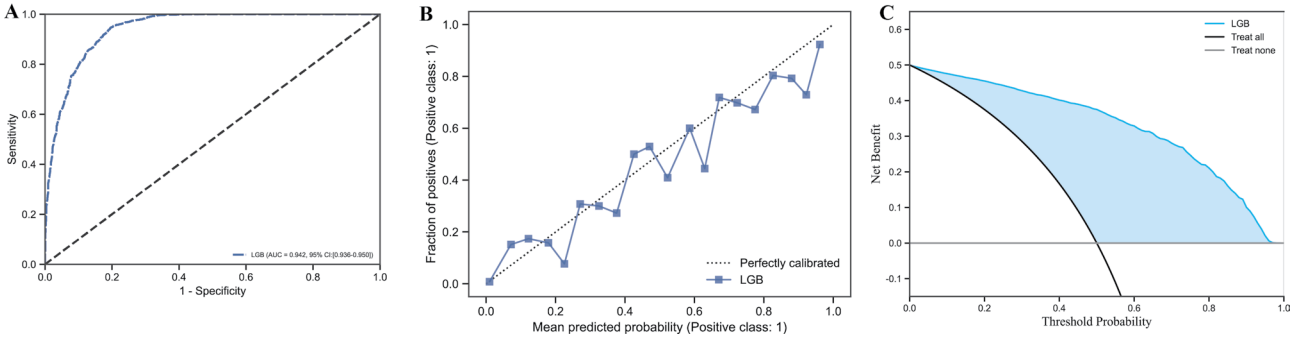

Fig. 7 Performance and clinical utility of the final LightGBM model in the external dataset. (**A**): Receiver operating characteristic (ROC) curve of the LGB model, plotting sensitivity against 1-specificity; (**B**): Calibration curve for the LGB model, comparing the mean predicted probability with the actual fraction of positive cases to assess prediction accuracy; (**C**): Decision curve analysis (DCA) for the LGB model, illustrating the net benefit of using the model for clinical decision-making across a range of threshold probabilities compared to treating all or no patients

Table 3 Comparison of discrimination performance across related works

Study	Dataset	Methodology	Performance	Study scope	Key contributions
Biswas et al. (2024) [27]	Diabetic foot ulcer (DFU) Dataset	XAI-FusionNet with multi-scale CNN fusion (DenseNet201, VGG19, NASNetMobile) and XAI (SHAP, LIME, Grad-CAM)	Accuracy: 99.05%, AUC: 99.09%	Binary classification of diabetic foot ulcers vs. healthy skin using images	Novel multi-scale fusion model with enhanced explainability for DFU detection
Biswas et al. (2024) [28]	Diabetic foot ulcer (DFU) Dataset	DFU_XAI: Compared DL models (e.g., ResNet50) with XAI (SHAP, LIME, Grad-CAM)	Accuracy: 98.75% (ResNet50)	Image-based binary DFU classification with model comparison	Robust DL model evaluation and XAI integration for DFU diagnosis
Biswas et al. (2023) [29]	Diabetic foot ulcer (DFU) Dataset	DFU_MultiNet: A hybrid deep learning model fusing features from DenseNet201, VGG19, & NasNetMobile in parallel.	Accuracy: 99.1%	Image-based binary classification of DFU vs. healthy skin.	- Proposes a novel hybrid, parallel deep learning architecture.
Kim et al. (2022) [30]	1581 patients (clinical & lab data)	Logistic regression for multi-class classification	AUC: 0.80 (necrotizing fasciitis), 0.73 (osteomyelitis)	Differentiation of diabetic foot, necrotizing fasciitis, and osteomyelitis using non-imaging data	Non-invasive model using routine data for early infection classification
Khandakar et al. (2022) [31]	167 foot-pair thermograms (Three severity classes)	K-means clustering + ML/DL comparison (e.g., VGG19, stacking classifier)	Accuracy: 95.08% (VGG19)	Severity classification of diabetic foot risk using thermogram images	Unsupervised labeling and ML/DL comparison for risk stratification
Khandakar et al. (2021) [32]	167 foot-pair thermograms (diabetic vs. control)	Compared CNNs (e.g., MobileNetV2) and ML (e.g., AdaBoost) on extracted features	F1 Score: 97% (AdaBoost)	Binary classification of diabetic vs. control feet using thermograms	Demonstrates ML superiority over CNNs for non-invasive detection
Reyes-Luévano et al. (2023) [33]	Multi-modal images (visible and infrared)	DFU_VIRNet: Dual-branch CNN for image processing	AUC: 0.9923 (DFU), 0.9982 (ischemia)	Classification of DFU, ischemia, and infection using multi-modal imaging	Novel CNN with risk zone visualization for early intervention
Das et al. (2022) [34]	1679 image patches (normal vs. abnormal)	DFU_SPNet: Stacked parallel CNN with multiple kernels	AUC: 0.974	Binary image-based DFU classification	Innovative CNN architecture with optimizer tuning for DFU detection
This Study	Two-center cohort: 3612 patients (DFI $n=1699$, OM $n=1913$)	LightGBM with XAI (SHAP, LIME) and feature selection	Internal AUC: 0.879, External AUC: 0.942	Differential diagnosis of DFI vs. OM using routine blood biomarkers	Externally validated, explainable ML model with web calculator for clinical use

Age was identified as the feature with the highest overall impact, where older age was strongly associated with a higher model prediction for osteomyelitis. This was followed in importance by HbA1c, creatinine, albumin (ALB), sodium (Na), and erythrocyte sedimentation rate (ESR). The plot further detailed that high values for HbA1c, creatinine, and ESR, along with low values for ALB and Na, consistently pushed the model's prediction higher. These global trends were confirmed by the SHAP heatmap, which visualized feature contributions across the entire dataset of instances sorted by prediction score (Fig. 9B). In this visualization, features such as Age consistently showed positive contributions (red coloring) for high-prediction instances and negative contributions (blue coloring) for low-prediction instances, a pattern also observed for HbA1c and creatinine. In a

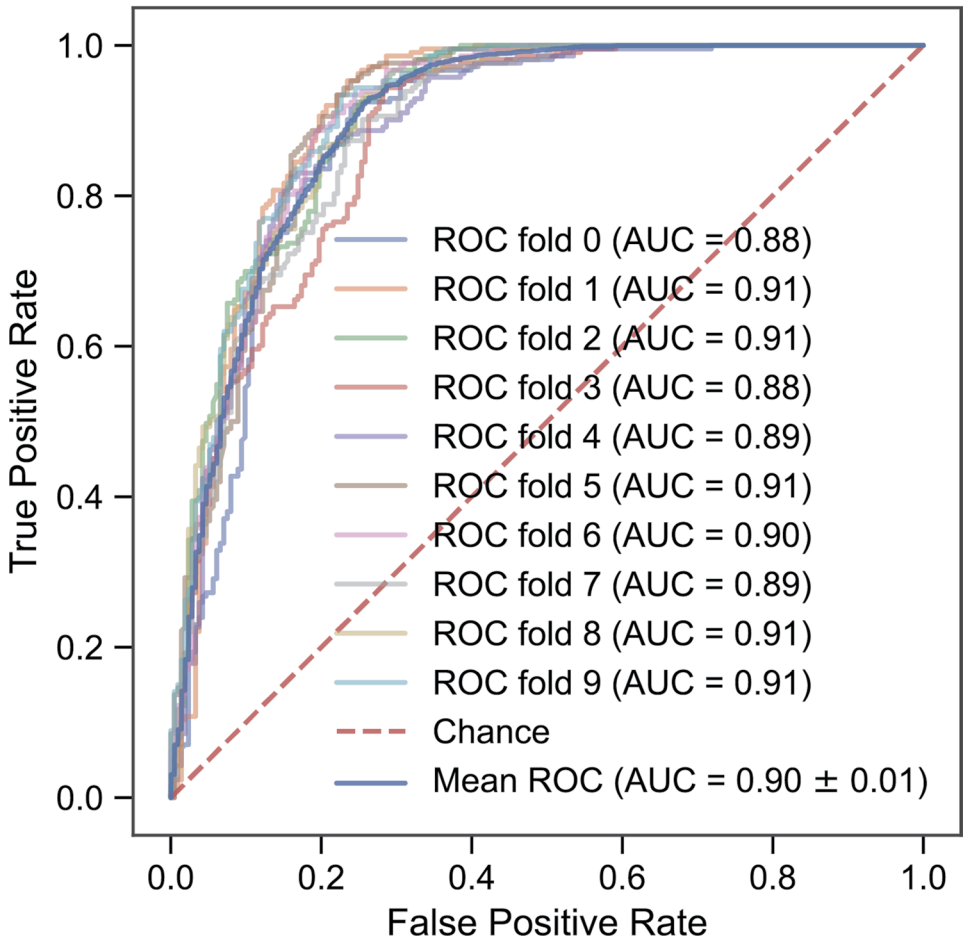

Fig. 8 Performance of the final model evaluated by 10-fold cross-validation on the overall dataset. The graph displays the receiver operating characteristic (ROC) curves from the 10-fold cross-validation. The plot presents the true positive rate (y-axis) versus the false positive rate (x-axis). The ten individual, semi-transparent lines show the performance for each of the validation folds. The solid dark blue line represents the mean ROC curve averaged from ten folds, and the dashed red line indicates the performance of a random chance classifier

complementary analysis, LIME was used to assess feature importance, which ranked HbA1c as the most influential feature, followed by Age, Na, Creatinine, ALB, and ESR (Fig. 9C). While the exact order differed slightly from SHAP, both methods consistently identified HbA1c and Age as the two most critical predictors. To understand the model's logic at an individual level, a waterfall plot for a single representative case was generated (Fig. 9D). This plot detailed how, for a specific instance, the model's base prediction (E[f(X)] = -0.645) was additively modified by the patient's specific feature values—such as an age of 58 (+1.15) and a creatinine level of 68.47 (+0.67)—to arrive at a final high-risk prediction score of 3.096.

To further explore the marginal effect of each feature on the model's prediction, we generated SHAP dependence plots for the six most influential biomarkers (Fig. 10). For albumin (ALB), an inverse relationship was observed; values below approximately 35 g/L corresponded to positive SHAP values, which increased the model's output, whereas higher ALB levels were associated with progressively negative SHAP values (Fig. 10A). In contrast, age exhibited a strong, positive monotonic relationship with the model output, where SHAP values became consistently positive for ages above approximately 40 years and increased steadily thereafter, indicating that older age was a significant driver of higher risk predictions (Fig. 10B). The plot for creatinine showed that while most values were concentrated at lower levels with a wide spread of SHAP values, extremely high values contributed positively to the model's prediction (Fig. 10C). The erythrocyte sedimentation rate (ESR) displayed a non-linear pattern, with a peak in positive SHAP values around 30–40 mm/h, while values outside this optimal range had a lesser or negative impact on the model's output (Fig. 10D). A clear positive correlation was evident for HbA1c, where values exceedingly approximately 2% consistently yielded positive SHAP values that increased in magnitude with the HbA1c level (Fig. 10E). Finally, the relationship for sodium (Na) was also non-linear; values centered around 125–135 mmol/L were linked to positive

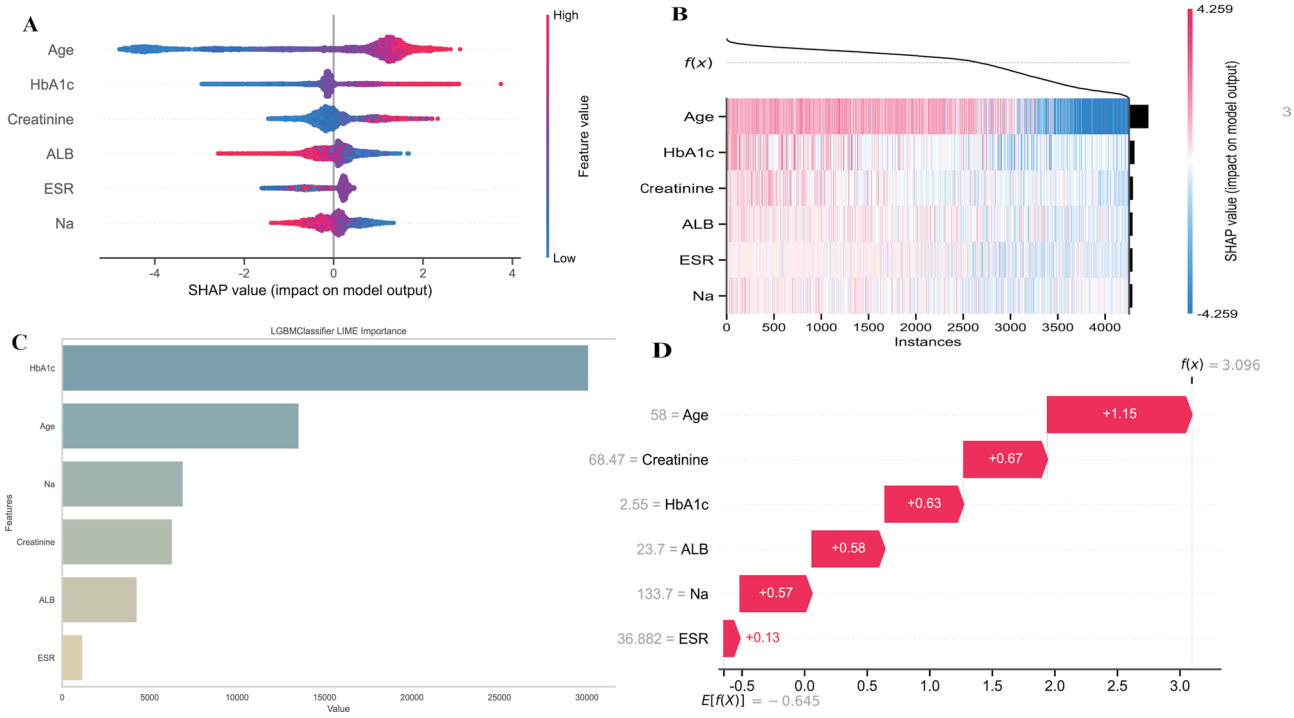

Fig. 9 Feature importance and model interpretability analysis. (**A**): A SHAP summary plot displaying features ranked by the sum of their SHAP value magnitudes across all samples. Each point on the plot represents a Shapley value for a feature and an instance, with the x-axis indicating the SHAP value's impact on the model output and color denoting the feature's value from low (blue) to high (red); (**B**): A SHAP heatmap visualizing the attributions of each feature across all instances. The x-axis represents individual instances sorted by similarity, the y-axis lists the features, and the color scale indicates the SHAP value, while the curve f(x) at the top shows the model's output for each corresponding instance; (**C**): A horizontal bar chart illustrating the global feature importance as determined by the LGBMClassifier LIME method, with features ranked along the y-axis and their calculated importance value shown on the x-axis; (**D**): A SHAP waterfall plot for a single prediction, illustrating how the positive (red) and negative (blue) contributions of each feature value summatively adjust the base model output (E[f(X)]) to yield the final prediction score (f(x))

SHAP values, whereas values deviating from this range, particularly higher concentrations above 135 mmol/L, corresponded to negative SHAP values (Fig. 10F).

Model online development

To facilitate clinical application and promote model transparency, the final LGB model was deployed as a user-friendly, interactive web application (https://df-osteo.streamlit.app/) (Fig. 11). This tool allows clinicians to input six routine patient biomarkers—Age (e.g., 58 years), Albumin (23.70 g/L), Creatinine (368.47 μmol/L), ESR (36.88 mm/hr), HbA1c (2.55%), and Sodium (133.70 mmol/L)—to receive an instant differential diagnosis. For the case presented, the tool predicted a diagnosis of Diabetic Foot Infection (DFI) with 95.0% confidence. To explain this prediction, the interface generated real-time SHAP force and waterfall plots. The DFI explanation showed how the combination of input features pushed the model's output from a base value of -0.645 to a final high-risk score of 3.096. The waterfall plot quantified these contributions, showing that Age (+1.15), Creatinine (+0.67), and HbA1c (+0.63) were the strongest drivers of the DFI diagnosis. Concurrently, the tool provided an explanation for the alternative diagnosis (Osteomyelitis), showing how the same feature values pushed the prediction score down to -2.096, with each feature providing an equal and opposite contribution (e.g., Age: -1.15), thereby providing a comprehensive and interpretable basis for the model's conclusion.

Discussion

The rising prevalence of diabetic foot complications in Africa was driven by rapid demographic and economic shifts, with diabetic peripheral neuropathy and peripheral arterial disease emerging as the primary etiologies. Regional amputation rates reached 61% in severe cases [35]. In Ethiopia, 12.98% of diabetic patients developed foot ulcers, exacerbated by rural residence and inadequate self-care practices [36]. Peruvian inpatient data revealed an 18.9% prevalence of diabetic foot pathology among diabetes patients, frequently involving osteomyelitis and severe infections [37]. Consistent risk factors across studies included advanced age, prolonged diabetes duration, and suboptimal glycemic control [2]. Neuropathy and ischemia were identified as the dominant pathological mechanisms precipitating ulceration and

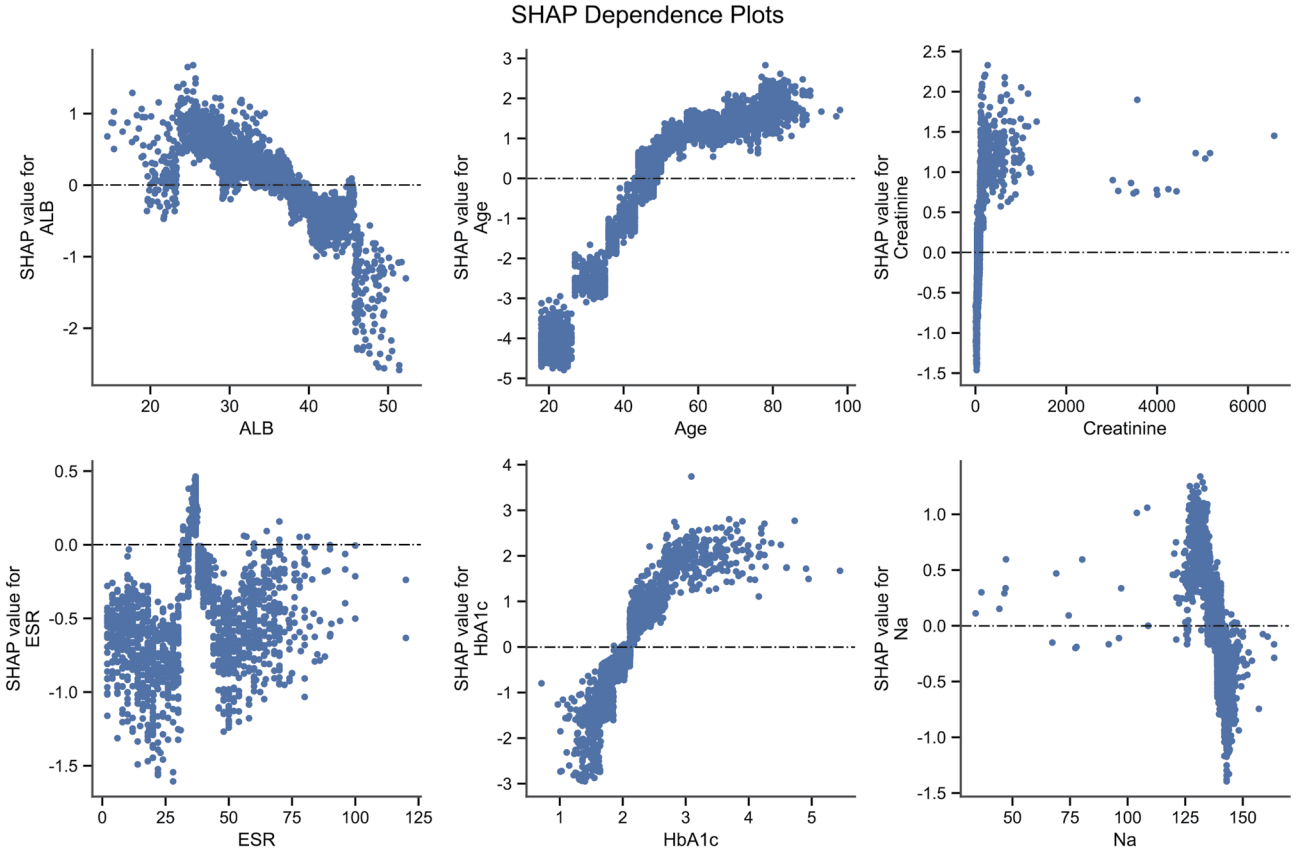

Fig. 10 SHAP dependence plots for the six most important features in the final model. (**A**): A scatter plot showing the relationship between albumin (ALB) feature values on the x-axis and their corresponding SHAP values on the y-axis, where each point represents a single observation from the dataset; (**B**): A dependence plot for Age, illustrating the feature's value on the x-axis against its SHAP value on the y-axis to show its marginal effect on the model output; (**C**): A plot depicting Creatinine values on the x-axis and their impact on the model prediction (SHAP value) on the y-axis; (**D**): A dependence plot for erythrocyte sedimentation rate (ESR), with feature values plotted on the x-axis and their corresponding SHAP values on the y-axis; (**E**): A plot showing the marginal effect of HbA1c, with its values on the x-axis and SHAP values on the y-axis; (**F**): A dependence plot for Sodium (Na), displaying the feature's values on the x-axis and their corresponding SHAP values on the y-axis

subsequent morbidity. Globally, diabetic foot disease affected 6.3% of patients, imposing significant quality-of-life and healthcare system burdens [38]. These findings underscored diabetic foot complications as a pressing public health challenge requiring integrated prevention and management approaches worldwide. Distinguishing OM from soft tissue infections (STIs) in DFIs remained a critical diagnostic challenge due to overlapping clinical presentations and limitations in conventional diagnostic methods.

In this two-center study, we addressed the critical and long-standing challenge of distinguishing diabetic foot infection (DFI) from osteomyelitis (OM) by developing and rigorously validating an interpretable machine learning model. Our final LightGBM model successfully leveraged just six routine, readily available blood biomarkers—Age, HbA1c, Creatinine, Albumin, ESR, and Sodium—to achieve high diagnostic accuracy. The model demonstrated robust performance not only in internal validation (AUC 0.879) but, more importantly, in an independent external cohort, where its discriminative power was excellent (AUC 0.942). By integrating explainable AI techniques and deploying the model as an interactive web calculator, we directly confronted the "black box" barrier that often hinders clinical adoption. To our knowledge, this work represents a significant step forward by creating a non-invasive, cost-effective, and transparent diagnostic tool that was validated across different clinical settings, offering a practical solution to an urgent unmet need in the management of diabetic foot complications.

Previous studies have demonstrated the potential of machine learning in addressing diabetic complications. Manjunath et al. examined data from 767 patients at an Indian medical center, employing algorithms like Random Forest, XGBoost, LightGBM, CatBoost, neural networks, and ensemble methods to predict diabetic retinopathy and nephropathy; they achieved an AUC of 1.0 on oversampled datasets and highlighted the benefits of model optimization for handling imbalanced

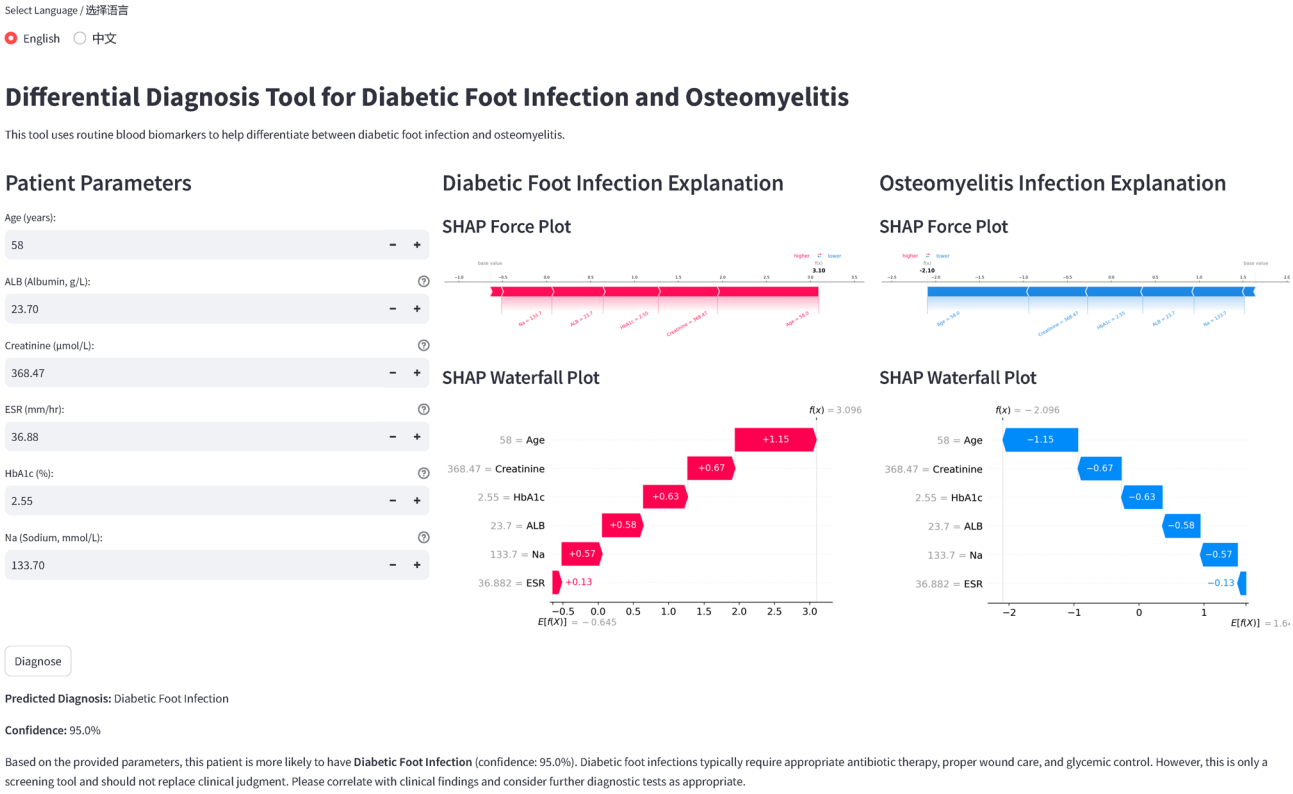

Fig. 11 User interface of the deployed web-based diagnostic tool for real-time prediction and interpretation. The screenshot displays the graphical user interface of the differential diagnosis tool, which is divided into several functional sections: the left panel contains input fields for patient parameters such as Age, Albumin, and Creatinine; below the input fields, a section displays the model's final predicted diagnosis and the associated confidence level; the central and right panels are dedicated to model explanation, providing SHAP force plots and SHAP waterfall plots that visualize how each patient parameter contributes to the likelihood of Diabetic Foot Infection versus Osteomyelitis, respectively

clinical data, suggesting integration into decision support systems for better personalized care and resource use [39]. Similarly, Dagliati et al. analyzed electronic health records from approximately 1,000 patients in the EU-funded MOSAIC project, using a data mining approach with random forest imputation and logistic regression to forecast retinopathy, neuropathy, and nephropathy at 3, 5, and 7 years, attaining up to 83.8% accuracy and developing models that are clinically applicable [40]. Guan et al. reviewed machine learning applications in diabetic foot care, emphasizing its role in improving diagnostic accuracy and treatment strategies through the use of medical imaging, biomarker analysis, and biomechanics, thereby transforming management of foot complications [41]. Oikonomou and Khera explored how machine learning facilitates personalized prediction and management of cardiovascular risks in diabetes, stressing the need for rigorous methods and bias reduction to ensure safe clinical adoption [42]. Collectively, these investigations illustrated that machine learning techniques, from ensemble classifiers to deep learning, effectively predict, diagnose, and guide treatment for diabetic complications, promising earlier interventions, customized care, and enhanced outcomes. However, they also underscored persistent challenges, such as the need for thorough model validation, greater data diversity, and ethical considerations, which our study addressed by incorporating explainable AI and external validation to advance reliable, translatable tools for clinical practice.

Previous research demonstrated the potential of machine learning in diagnosing osteomyelitis and related complications. JiYeol Kim et al. analyzed data from 1,581 patients with diabetic foot ulcers, including cases of osteomyelitis and necrotizing fasciitis, to develop a classification model using demographic and routine laboratory data. They applied feature selection and various algorithms, achieving AUCs of 0.80 for necrotizing fasciitis and 0.73 for osteomyelitis in validation sets. These results highlighted the model's ability to support accurate, non-invasive diagnosis and facilitate timely

treatment decisions without relying on advanced testing [30]. Similarly, Sun-Gyu Choi et al. studied 222 patients with jaw osteomyelitis, utilizing logistic regression, random forest, support vector machine, artificial neural networks, and extreme gradient boosting. Their models achieved AUCs ranging from 0.81 to 0.88, with ensemble methods proving superior in identifying key risk factors that could guide prognosis and treatment strategies, offering insights adaptable to other osteomyelitis cases [43]. These studies illustrated how machine learning, when based on accessible clinical data, enhances early detection and management of osteomyelitis, improving patient outcomes and resource efficiency. Our study extended this foundation by incorporating explainable AI and external validation, addressing gaps in generalizability and interpretability. This approach provided a more robust tool for managing diabetic foot complications, paving the way for broader clinical adoption and better-informed decision-making in everyday practice. The integration of a clinically applicable web calculator further underscored the practical utility of our findings, aligning with the demand for transparent and actionable solutions in medical diagnostics.

Aroa et al. conducted a prospective study of 116 diabetic foot osteomyelitis patients and showed that higher albumin and lymphocyte counts predicted better 12-month outcomes, whereas elevated ESR, CRP and glycemia were diagnostic for complications, supporting serial blood monitoring for early detection and prevention of adverse events [44]. Guojun Guo et al. used a cohort of 23,434 diabetic patients from the UK Biobank in a prospective analysis and applied Cox models with Mendelian randomization to demonstrate that higher HbA1c levels were strongly linked to increased risk of lower-limb ulcers, with risk rising steeply above 53 mmol/mol, suggesting that maintaining HbA1c below this threshold could reduce ulcer incidence [45]. Seyed Kaveh Moallemi et al. evaluated 142 diabetic foot patients in a cross-sectional study and found that ESR (cut-off 49 mm/h) had fair diagnostic accuracy (AUC 0.70) for osteomyelitis, whereas CRP showed poor performance, indicating that ESR may serve as a moderately useful marker in clinical practice [46]. J. Aragón-Sánchez et al. performed a prospective observational study on older adults with diabetes-related foot infections and reported that advanced age, higher creatinine and lower albumin were significant predictors of minor and major amputations, underscoring the influence of renal function and nutritional status on patient outcomes [47]. These showed that albumin consistently emerged as a protective factor, with higher levels linked to better healing and reduced amputation risk, while age and creatinine reflected comorbidity burden and renal impairment that compounded adverse outcomes. ESR served as a moderate diagnostic and prognostic marker for osteomyelitis, whereas HbA1c demonstrated robust predictive value for ulcer development and complications, reinforcing the necessity of tight glycemic control. Although sodium received limited attention, the overall evidence supported the clinical utility of routinely measured biomarkers—albumin, age, creatinine, ESR and HbA1c—for early detection, which were selected most powerful features, risk assessment and personalized management of diabetic foot and osteomyelitis complications, thereby enabling timely interventions, improving limb salvage rates and reducing morbidity.

Our study demonstrated that an explainable machine learning model using routine blood biomarkers effectively differentiated diabetic foot infection from osteomyelitis, with key features such as age, HbA1c, creatinine, albumin, ESR, and sodium emerging as critical predictors through SHAP analysis, achieving high AUC values and robust external validation. This aligns with previous research by Senneville et al. [48] and Berendt et al. [49], who identified similar biomarkers like inflammatory markers and age as influential in diagnosing diabetic foot osteomyelitis, likely due to their reflection of systemic inflammation and metabolic dysregulation, which our findings corroborated through consistent feature importance rankings and strong discriminative performance. However, inconsistencies arose with studies that incorporated imaging data or broader clinical variables, potentially leading to higher predictive accuracy in their models; for instance, The other studies reported enhanced outcomes with multimodal data, which our biomarker-only approach may not fully capture, possibly because of differences in data richness or patient cohorts, such as varying prevalences of comorbidities, highlighting the trade-off between model simplicity and comprehensiveness in real-world applicability [50, 51].

In contrast to these earlier works, our focus on routine, cost-effective biomarkers and explainable AI techniques addressed gaps in generalizability and interpretability, as evidenced by our model's stable performance across validations, though we recognize potential limitations in scenarios with limited data diversity. While Senneville et al. [48] emphasized the role of imaging in complex cases, our results suggest that a biomarker-driven model can still provide reliable insights, potentially due to standardized laboratory measures reducing variability, yet this deviation underscores the need for hybrid approaches in future studies to integrate multiple data sources. Overall, by building on these foundations, our work advances the field by offering a clinically deployable tool that enhances diagnostic precision in resource-constrained settings, though challenges in broader validation persist, urging cautious adoption and further research to refine such models for optimal patient care.

Limitations

Despite its robust design and promising results, this study has several important limitations. First and foremost, a primary limitation is its binary classification framework, which exclusively focuses on differentiating DFI from OM. This approach, while targeting a specific and challenging clinical question, does not represent the full, heterogeneous spectrum of diabetic foot pathologies, which includes non-infected ischemic or neurotrophic ulcers, Charcot neuroarthropathy, and other inflammatory conditions. Consequently, the model's high-performance metrics may be partly attributable to this constrained diagnostic context. It is crucial to emphasize that the developed model is intended as a decision support tool for cases where infection is already strongly suspected, not as a general-purpose screening tool for all diabetic foot complications. Its application outside this specific DFI-versus-OM differential diagnosis is inappropriate and could yield misleading results. Future work is essential to develop and validate multi-class models that incorporate "other" or "non-infectious" category to improve clinical generalizability. Second, the study's retrospective nature introduces an inherent risk of selection bias and prevents the establishment of causality. Furthermore, our reliance on a composite reference standard, though reflective of real-world clinical practice, is less definitive than the universal application of bone biopsy and histopathology, which could introduce a degree of label noise into the dataset. Third, while external validation across two centers is a key strength, both institutions are located within the same geographic region of China. The model's performance thus requires further validation across more ethnically and geographically diverse populations to ensure broader applicability. Finally, our model is based on a static snapshot of biomarkers from a single time point upon admission. It does not capture the temporal dynamics of these markers, which could offer additional diagnostic or prognostic information. The model relies solely on routinely available blood biomarkers. Important diagnostic information from clinical examination, imaging (radiography, MRI), microbiology, and wound characteristics were not included in the model. Integrating multimodal data may be required to achieve clinically actionable performance in broader use cases. Future prospective studies incorporating longitudinal data and other modalities are needed to address these limitations and further validate the model's real-world clinical utility.

Conclusions

In conclusion, this study developed and externally validated an interpretable machine learning model capable of distinguishing DFI from OM using routine blood biomarkers. The final LightGBM model demonstrated high discriminative performance in both internal and external validation cohorts, was well-calibrated, and showed positive net benefit in decision curve analysis. The integration of explainable AI techniques, such as SHAP, provides transparency into the model's predictions, and the deployment of this model as a publicly accessible web calculator offers a practical means for clinical translation. This work presents a non-invasive and low-cost tool that, by leveraging readily available data, has the potential to aid clinicians in the diagnostic process for complex diabetic foot cases, particularly in resource-limited settings.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03236-9.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

P.Y. and S.D. contributed equally to this work. P.Y., S.D., M.Y., and X.S. conceptualized and designed the study. P.Y., S.D., Z.A., and A.Y. were responsible for data acquisition and curation. P.Y. and S.D. conducted formal analysis, developed machine learning models, and implemented software. Y.Y. contributed to data validation and investigation. All authors participated in the interpretation of the results. P.Y. and S.D. wrote the original manuscript draft. M.Y. and X.S. supervised the project, provided resources, and critically reviewed and edited the final manuscript. All authors have read and approved the final version for submission.

Funding

This work was financially supported by the following grants: the Health Care and Medical Research Special Project of the Xinjiang Uygur Autonomous Region (Grant No. BL202460); the Youth Medical Science and Technology Talent Special Research Project (Grant No. WJWY-202331); the National Natural Science Foundation of China (Grant No. 82460420); the Research and Innovation Team Project of Xinjiang Medical University (Grant No. XYD2024C12); the Tianshan Talent Training Program (Grant No. TSYC202301B039); the Key Laboratory of High Incidence Disease Research in Xinjiang (Xinjiang Medical University), Ministry of Education (Grant No. 2023B04); and the Natural Science Foundation of the Xinjiang Uygur Autonomous Region (Grant No. 2022D01C821). These funding sources provided critical resources for the successful execution of this research.

Data availability

Upon a reasonable request, the corresponding authors of this article will provide unrestricted access to the original data.

Declarations

Ethics approval and consent to participate

This study adhered to the principles outlined in the Helsinki Declaration and received approval from the Ethics Committee of The Sixth Affiliated Hospital of Xinjiang Medical University and First People's Hospital of Kashi Prefecture. The requirement for informed consent was waived by the Ethical Committee, as the study involved de-identified data, posing no potential risk to patients and maintaining no connection between the patients and researchers. All procedures were conducted in compliance with the applicable guidelines and regulations. No biological specimens were used in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details
[1]Department of Spine Surgery, The Sixth Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830000, People's Republic of China
[2]Department of Spine Surgery, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830000, People's Republic of China
[3]The First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang 830000, People's Republic of China
[4]Xinjiang Key Laboratory of Artificial Intelligence Assisted Imaging Diagnosis, Department of Radiology, The First People's Hospital of Kashi Prefecture, Kashi, Xinjiang 844000, People's Republic of China

Received: 7 August 2025 / Accepted: 6 October 2025
Published online: 11 November 2025

References

1. McDermott K, Fang M, Boulton AJM, Selvin E, Hicks CW. Etiology, Epidemiology, and disparities in the burden of diabetic foot ulcers. Diabetes Care. 2023;46(1):209–21.
2. Fang M, Hu J, Jeon Y, Matsushita K, Selvin E, Hicks CW. Diabetic foot disease and the risk of major clinical outcomes. Diabetes Res Clin Pract. 2023;202:110778.
3. Stancu B, Ilyes T, Farcas M, Coman HF, Chis BA, Andercou OA. Diabetic foot complications: a retrospective cohort study. Int J Environ Res Public Health. 2022;20(1).
4. Edmonds M, Manu C, Vas P. The current burden of diabetic foot disease. J Clin Orthop Trauma. 2021;17:88–93.
5. Sen P, Demirdal T. Evaluation of mortality risk factors in diabetic foot infections. Int Wound J. 2020;17(4):880–9.
6. Korpinar S. A retrospective analysis of microbiologic profile of foot infections in patients with diabetic End-Stage renal disease. Int J Low Extrem Wounds. 2021;20(1):15–21.
7. Lauri C, Leone A, Cavallini M, Signore A, Giurato L, Uccioli L. Diabetic foot infections: the diagnostic challenges. J Clin Med. 2020;9(6).
8. Senneville EM, Lipsky BA, van Asten SAV, Peters EJ. Diagnosing diabetic foot osteomyelitis. Diabetes Metab Res Rev. 2020;36(Suppl 1):e3250.
9. Mponponsuo K, Sibbald RG, Somayaji R. A comprehensive review of the Pathogenesis, Diagnosis, and management of diabetic foot infections. Adv Skin Wound Care. 2021;34(11):574–81.
10. Carro GV, de Jesus FM, Ricci A. Diabetic foot osteomyelitis: is it all the same? Int J Low Extrem Wounds. 2023:15347346231160614.
11. Rubitschung K, Sherwood A, Crisologo AP, Bhavan K, Haley RW, Wukich DK, Castellino L, Hwang H, La Fontaine J, Chhabra A, et al. Pathophysiology and molecular imaging of diabetic foot infections. Int J Mol Sci. 2021;22(21).
12. Glaudemans AW, Uckay I, Lipsky BA. Challenges in diagnosing infection in the diabetic foot. Diabet Med. 2015;32(6):748–59.
13. Llewellyn A, Kraft J, Holton C, Harden M, Simmonds M. Imaging for detection of osteomyelitis in people with diabetic foot ulcers: A systematic review and meta-analysis. Eur J Radiol. 2020;131:109215.
14. Lauri C, Tamminga M, Glaudemans A, Juarez Orozco LE, Erba PA, Jutte PC, Lipsky BA, MJ IJ, Signore A, Slart R. Detection of osteomyelitis in the diabetic foot by imaging techniques: A systematic review and Meta-analysis comparing MRI, white blood cell Scintigraphy, and FDG-PET. Diabetes Care. 2017;40(8):1111–20.
15. Lam K, van Asten SA, Nguyen T, La Fontaine J, Lavery LA. Diagnostic accuracy of probe to bone to detect osteomyelitis in the diabetic foot: A systematic review. Clin Infect Dis. 2016;63(7):944–8.
16. Crisologo PA, Davis KE, Ahn J, Farrar D, Van Asten S, La Fontaine J, Lavery LA. The infected diabetic foot: can serum biomarkers predict osteomyelitis after hospital discharge for diabetic foot infections? Wound Repair Regen. 2020;28(5):617–22.
17. Ansert EA, Tarricone AN, Coye TL, Crisologo PA, Truong D, Suludere MA, Lavery LA. Update of biomarkers to diagnose diabetic foot osteomyelitis: A meta-analysis and systematic review. Wound Repair Regen. 2024;32(4):366–76.
18. Yammine K, Abou Orm G, Mouawad J, Assi C. Basic haematological tests as inflammatory performance markers of patients treated either by Conservative surgery or minor amputation for infected diabetic foot ulcers. Wound Repair Regen. 2023;31(5):627–34.
19. Khanna NN, Maindarkar MA, Viswanathan V, Puvvula A, Paul S, Bhagawati M, Ahluwalia P, Ruzsa Z, Sharma A, Kolluri R, et al. Cardiovascular/stroke risk stratification in diabetic foot infection patients using deep learning-based artificial intelligence: an investigative study. J Clin Med. 2022;11(22).
20. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. J Gastroenterol Hepatol. 2021;36(3):581–4.
21. Senneville E, Albalawi Z, van Asten SA, Abbas ZG, Allison G, Aragon-Sanchez J, Embil JM, Lavery LA, Alhasan M, Oz O, et al. IWGDF/IDSA guidelines on the diagnosis and treatment of diabetes-related foot infections (IWGDF/IDSA 2023). Diabetes Metab Res Rev. 2024;40(3):e3687.
22. Bonnet E, Maulin L, Senneville E, Castan B, Fourcade C, Loubet P, Poitrenaud D, Schuldiner S, Sotto A, Lavigne JP, et al. Clinical practice recommendations for infectious disease management of diabetic foot infection (DFI) – 2023 SPILF. Infect Dis now. 2024;54(1):104832.
23. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinf. 2019;20(1):492.
24. Li R, Shinde A, Liu A, Glaser S, Lyou Y, Yuh B, Wong J, Amini A. Machine Learning-Based interpretation and visualization of nonlinear interactions in prostate cancer survival. JCO Clin Cancer Inf. 2020;4:637–46.
25. Patterson J, Tatonetti N. KG-LIME: predicting individualized risk of adverse drug events for multiple sclerosis disease-modifying therapy. J Am Med Inf Assoc. 2024;31(8):1693–703.
26. Parker A, Heflin A, Jones LC. Analyzing university of Virginia health publications using open data, Python, and streamlit. J Med Libr Assoc. 2021;109(4):688–9.
27. Biswas S, Mostafiz R, Uddin MS, Paul BK. XAI-FusionNet: diabetic foot ulcer detection based on multi-scale feature fusion with explainable artificial intelligence. Heliyon. 2024;10(10):e31228.
28. Biswas S, Mostafiz R, Paul BK, Uddin KMM, Hadi MA, Khanom F. DFU_XAI: A deep Learning-Based approach to diabetic foot ulcer detection using feature explainability. Biomed Mater Devices. 2024;2(2):1225–45.
29. Biswas S, Mostafiz R, Paul BK, Mohi Uddin KM, Rahman MM, Shariful FNU. DFU_MultiNet: A deep neural network approach for detecting diabetic foot ulcers through multi-scale feature fusion using the DFU dataset. Intelligence-Based Med. 2023;8:100128.
30. Kim J, Yoo G, Lee T, Kim JH, Seo DM, Kim J. Classification model for diabetic foot, necrotizing fasciitis, and osteomyelitis. Biology (Basel). 2022;11(9).
31. Khandakar A, Chowdhury MEH, Reaz MBI, Ali SHM, Kiranyaz S, Rahman T, Chowdhury MH, Ayari MA, Alfkey R, Bakar AAA, et al. A novel machine learning approach for severity classification of diabetic foot complications using thermogram images. Sens (Basel). 2022;22(11).
32. Khandakar A, Chowdhury MEH, Ibne Reaz MB, Md Ali SH, Hasan MA, Kiranyaz S, Rahman T, Alfkey R, Bakar AAA, Malik RA. A machine learning model for early detection of diabetic foot using thermogram images. Comput Biol Med. 2021;137:104838.
33. Reyes-Luévano J, Guerrero-Viramontes JA, Romo-Andrade JR, Funes-Gallanzi M. DFU_VIRNet: A novel Visible-InfraRed CNN to improve diabetic foot ulcer classification and early detection of ulcer risk zones. Biomed Signal Process Control. 2023;86.
34. Das SK, Roy P, Mishra AK. DFU_SPNet: a stacked parallel convolution layers based CNN to improve diabetic foot ulcer classification. ICT Express. 2022;8(2):271–5.
35. Abbas ZG, Boulton AJM. Diabetic foot ulcer disease in African continent: 'From clinical care to implementation'- Review of diabetic foot in last 60 years – 1960 to 2020. Diabetes Res Clin Pract. 2022;183:109155.
36. Tolossa T, Mengist B, Mulisa D, Fetensa G, Turi E, Abajobir A. Prevalence and associated factors of foot ulcer among diabetic patients in ethiopia: a systematic review and meta-analysis. BMC Public Health. 2020;20(1):41.
37. Yovera-Aldana M, Saenz-Bustamante S, Quispe-Landeo Y, Aguero-Zamora R, Salcedo J, Sarria C, Gonzales-Grandez N, Briceno-Alvarado M, Antezana-Roman A, Manrique H, et al. Nationwide prevalence and clinical characteristics of inpatient diabetic foot complications: A Peruvian multicenter study. Prim Care Diabetes. 2021;15(3):480–7.
38. Craus S, Mula A, Coppini DV. The foot in diabetes - a reminder of an ever-present risk. Clin Med (Lond). 2023;23(3):228–33.
39. Manjunath DR, Lohith JJ, Kumar SS, Das A. Predicting diabetic retinopathy and nephropathy complications using machine learning techniques. IEEE Access. 2025;13:70228–53.
40. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, De Cata P, Chiovato L, Bellazzi R. Machine learning methods to predict diabetes complications. J Diabetes Sci Technol. 2018;12(2):295–302.

41. Guan HF, Wang Y, Niu P, Zhang YX, Zhang YJ, Miao RY, Fang XY, Yin RY, Zhao S, Liu J, et al. The role of machine learning in advancing diabetic foot: a review. Front Endocrinol. 2024;15.
42. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol. 2023;22(1).
43. Choi SG, Lee EY, Lee OJ, Kim S, Kang JY, Lim JS. Prediction models for early diagnosis of actinomycotic osteomyelitis of the jaw using machine learning techniques: a preliminary study. BMC Oral Health. 2022;22(1).
44. Tardaguila-García A, Alvarez YG, García-Morales E, Alvaro-Afonso FJ, Sanz-Corbalán I, Lázaro-Martínez JL. Utility of blood parameters to detect complications during long-term follow-up in patients with diabetic foot osteomyelitis. J Clin Med. 2020;9(11).
45. Guo GJ, Guan YL, Chen YH, Ye YG, Gan ZY, Cao X, Chen ZB, Hao XJ. HbA1c and the risk of lower limb ulcers among diabetic patients: an observational and genetics study. J Diabetes Res. 2025;2025(1).
46. Moallemi SK, Niroomand M, Tadayon N, Forouzanfar MM, Fatemi A. Diagnostic value of erythrocyte sedimentation rate and C reactive protein in detecting diabetic foot Osteomyelitis; a Cross-sectional study. Arch Acad Emerg Med. 2020;8(1):e71.
47. Aragon-Sanchez J, Viquez-Molina G, Lopez-Valverde ME, Aragon-Hernandez C, Aragon-Hernandez J, Rojas-Bonilla JM. Clinical Features, inflammatory Markers, and limb salvage in older adults with Diabetes-Related foot infections. Int J Low Extrem Wounds. 2025;24(1):212–8.
48. Senneville É, Lipsky B, Van Asten S, Peters E. Diagnosing diabetic foot osteomyelitis. Diab/Metab Res Rev. 2020;36.
49. Berendt AR, Peters EJ, Bakker K, Embil JM, Eneroth M, Hinchliffe RJ, Jeffcoate WJ, Lipsky BA, Senneville E, Teh J, et al. Diabetic foot osteomyelitis: a progress report on diagnosis and a systematic review of treatment. Diabetes Metab Res Rev. 2008;24(Suppl 1):S145–161.
50. Tardáguila-García A, Sanz-Corbalán I, García-Alamino JM, Ahluwalia R, Uccioli L, Lázaro-Martínez JL. Medical versus surgical treatment for the management of diabetic foot osteomyelitis: a systematic review. J Clin Med. 2021;10(6).
51. Da Ros R, Assaloni R, Michelli A, Brunato B, Miranda C. Antibiotic and surgical treatment of diabetic foot osteomyelitis: the histopathological evidence. Antibiotics-Basel. 2024;13(12).

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.