



Machine learning in diabetic-foot osteomyelitis: current applications, challenges, and future directions

Farzad Pourghazi¹, Seyed Mohammad Amin Alavi², Francesco Petri^{1,3}, Takahiro Matsuo⁴, Sergio Alvarez Mulett¹, Fabio Borgonovo^{1,3}, Andrea Gori^{3,5,6}, Michael E. Torchia⁷, Aaron J. Tande¹, and Elie F. Berbari¹

¹Division of Public Health, Infectious Diseases and Occupational Medicine, Department of Medicine, Mayo Clinic College of Medicine and Science, Mayo Clinic, Rochester, MN, USA

²Faculty of Medicine, Ahvaz Jundishapur University of Medical Sciences, 6135715753 Ahvaz, Iran

³Department of Infectious Diseases, Luigi Sacco University Hospital, Regional Center for Infectious Diseases (CEREMI), 20157 Milan, Italy

⁴Department of Infectious Diseases, Infection Control, and Employee Health, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

⁵Department of Biomedical and Clinical Sciences, University of Milan, 20157 Milan, Italy

⁶Centre for Multidisciplinary Research in Health Science (MACH), University of Milan, 20157 Milan, Italy

⁷Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA

Correspondence: Elie F. Berbari (berbari.elie@mayo.edu)

Received: 18 August 2026 – Revised: 26 September 2026 – Accepted: 29 September 2026 – Published: 13 October 2026

Abstract. Background: Diabetic-foot osteomyelitis (DFO) is a severe complication of diabetic-foot ulcers associated with delayed healing, amputation, and increased mortality. Machine learning (ML) has increasingly been explored for diagnosis and clinical decision-making in diabetic-foot care; however, the current evidence specific to DFO has not been systematically synthesized. **Methods:** We performed systematic searches of MEDLINE, Embase, Scopus, and Web of Science for studies published between January 2000 and May 2026. Studies that developed or evaluated ML models for DFO diagnosis, prediction, or outcomes were included. Data extraction was performed using the CHARMS framework, reporting followed PRISMA 2020 and TRIPOD-SRMA recommendations, and methodological quality was evaluated using PROBAST + AI. Due to considerable heterogeneity, narrative synthesis was performed. **Results:** A total of 10 studies published from 2022 to 2026 met the inclusion criteria. Five studies developed or evaluated ML models with a direct DFO diagnostic endpoint, one additional study evaluated model explainability, and four evaluated complementary DFO-relevant diabetic-foot applications. Reported discrimination ranged from 0.879 to 0.998, but comparisons were limited by differences in clinical endpoints, reference standards, input modalities and validation strategies. Only one direct DFO model was externally validated, and no study developed a multimodal model combining clinical and imaging data. **Conclusion:** Current evidence is limited by heterogeneous methodologies, limited external validation, and inconsistent reporting despite promising discrimination reported by several models.

Future studies should evaluate whether multimodal models integrating clinical and imaging data provide incremental value over unimodal approaches, with external validation, calibration, and clinical utility assessment before clinical implementation.

1 Introduction

Diabetic-foot ulcers (DFUs) are common and disabling complications of diabetes. Globally, an estimated 529 million adults had diabetes in 2021 (Ong et al., 2023), and 19 %–34 % will develop a DFU during their lifetime (McDermott et al., 2023). More than half of DFUs become infected, and approximately 20 % of moderate-to-severe diabetic-foot infections result in amputation (Armstrong et al., 2023). DFU is also associated with longer lengths of stay, greater emergency department utilization, and 2.5-fold-higher 5-year mortality than diabetes without DFU (Armstrong et al., 2023). The American College of Radiology reports that 20 % of patients with DFU will develop osteomyelitis (Baker et al., 2025).

Beyond amputation risk, diabetic-foot osteomyelitis (DFO) imposes substantial clinical and economic burdens, often requiring prolonged antibiotics, repeated testing and procedures, hospitalization, and multidisciplinary care (Giurato et al., 2017). The 2023 International Working Group on the Diabetic Foot/Infectious Diseases Society of America (IWGDF/IDSA) guideline recommends considering surgical resection of infected bone combined with systemic antibiotics in selected patients with DFO (Senneville et al., 2023), highlighting the importance of accurate diagnosis for treatment and limb preservation decisions.

Diagnosing DFO is challenging as no single test is accurate enough and because its findings overlap with other conditions, including Charcot neuroarthropathy, trauma, and soft-tissue infection (Senneville et al., 2023).

The IWGDF/IDSA guideline recommends initial studies including clinical assessment with probe-to-bone testing, plain radiography, and inflammatory biomarkers. Magnetic resonance imaging (MRI) is indicated if the diagnosis is uncertain and bone biopsy can be considered for microbiological or histopathological confirmation (Senneville et al., 2023).

Plain radiography has limited sensitivity (61.9 %) but moderate specificity (78.3 %) for DFO (Llewellyn et al., 2020), whereas serial imaging may improve sensitivity but reduce specificity (Harrison et al., 2026). MRI has higher diagnostic accuracy, with a reported sensitivity and specificity of 96.4 % and 83.8 %, respectively, although there is substantial variability across studies (Llewellyn et al., 2020). The IWGDF/IDSA rates the evidence for advanced imaging in DFO as moderate because of inconsistency, imprecision, and indirectness (Senneville et al., 2023), highlighting the potential value of integrating complementary variables rather than relying on a single test (Llewellyn et al., 2020).

Recently, machine learning (ML) has been increasingly explored to support diagnostic classification and outcome prediction by modeling complex relationships among clinical, laboratory, and imaging variables. In diabetic-foot care, ML applications have incorporated diverse data sources, in-

cluding medical imaging, biomarkers, and clinical information, to support diagnostic and clinical decision-making (Guan et al., 2024; Misir, 2025). Across DFO studies, however, the predictors and features used have varied according to the data modality and modeling approach.

Clinical decision-making involves integrating symptoms, wound findings, laboratory results, and vascular status and imaging (Kline et al., 2022). Multimodal ML can integrate these complementary data sources, but the value of this approach should be demonstrated by direct comparison with unimodal models in the same population and appropriate validation rather than being assumed due to model complexity (Kline et al., 2022).

Recent reviews of machine learning in diabetic-foot care have described applications involving thermography, ulcer photography, structured clinical data, sensors, risk prediction, and outcome classification, but these reviews were not focused on DFO-specific clinical endpoints and reference standards (Guan et al., 2024; Misir, 2025). Therefore, this systematic review summarizes the evidence regarding ML for DFO diagnosis, risk prediction, and treatment outcomes, focusing on data sources, clinical endpoints, reference standards, model performance, validation methods, and gaps in the evidence.

2 Materials and methods

2.1 Review design and reporting

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement and the TRIPOD-SRMA guideline for systematic reviews of prediction model studies (Page et al., 2021; Snell et al., 2023). The review questions and extraction framework were informed by the CHARMS checklist (Moons et al., 2014). Relevant reporting considerations from TRIPOD+AI were also used to characterize machine-learning and deep-learning studies (Moons et al., 2025; Collins et al., 2024b). An internal review protocol was prepared before study initiation. The review was not registered, and the protocol is not publicly available; however, it can be provided by the corresponding author upon reasonable request.

2.2 Eligibility criteria

We included original human studies evaluating, developing, or interpreting ML or artificial intelligence (AI) models relevant to DFO. Eligible endpoints were DFO detection, differentiation from other diabetic-foot conditions, lesion localization, future DFO development, and clinical outcomes in patients with established DFO.

Diabetic-foot studies that complemented the DFO model by using osteomyelitis as a predictor, severity component, or target of model interpretability were included as these

could inform candidate predictors and future DFO models. Demographic, clinical, wound, vascular, laboratory, photographic, radiographic, MRI, and radiomic data were eligible inputs. Studies had to develop, evaluate, or interpret an eligible ML and/or AI model and report a quantitative performance measure. Studies using conventional statistical models only were excluded unless used as comparators. Exclusion criteria were reviews, meta-analyses, editorials, letters, case reports, animal studies, non-peer-reviewed reports without full text, studies published before 2000, and studies without an eligible ML and/or AI model.

2.3 Information sources and search strategy

A comprehensive search was conducted using Ovid MEDLINE, Ovid Embase, Scopus, and Web of Science from 1 January 2000 to 15 May 2026 using controlled vocabulary and free-text terms for diabetic-foot/ulcer, osteomyelitis/bone infection, AI/ML, prediction/classification, diagnosis, validation, model performance, and treatment outcomes. Searches were limited to English-language human studies. Full database-specific strategies are provided in the Supplement.

2.4 Study selection

Records were imported into Covidence (Veritas Health Innovation, Melbourne, Australia), and duplicates were removed. Two reviewers (FP and SMAA) independently screened titles and/or abstracts and assessed potentially eligible full texts. Disagreements were resolved by discussion, and full-text exclusion reasons were recorded in the PRISMA flow diagram.

2.5 Data extraction

A standardized extraction form based on the CHARMS was piloted and refined. Two reviewers (FP and SMAA) independently extracted and checked study characteristics, population, clinical endpoints, intended use, prediction time point, reference standards, predictors, imaging characteristics, preprocessing, feature selection, algorithms, model development and validation methods, discrimination, calibration, classification and localization metrics, and explainability methods.

When studies evaluated multiple algorithms, all eligible models were extracted. For the main tables, we presented the authors' primary or final model; when none was specified, we selected the model with the most complete clinically relevant validation results. Development and internal- and external-validation results were extracted separately.

2.6 Model endpoint classification

Synthesis was aggregated by clinical endpoint rather than algorithm type. Studies were categorized into (1) binary DFO detection and/or classification; (2) multiclass or differential

diagnosis with DFO; (3) DFO localization or explainability; and (4) complementary diabetic-foot models with osteomyelitis as a predictor, severity component, or model context. This approach allowed for the comparison of models addressing similar clinical questions while avoiding comparisons across heterogeneous model types.

2.7 Risk of bias, quality, and applicability

Risks of bias, methodological quality, and applicability were assessed using PROBAST + AI across four domains: participants/data sources, predictors, outcome, and analysis. Two reviewers (FP and SMAA) independently assessed signalling questions and reached consensus; pre-consensus assessments were retained, and inter-rater agreement was measured using percentage agreement and Cohen's kappa. The disagreements and their type were recorded. For explainability-oriented studies, PROBAST + AI was applied to the underlying prediction model, and explainability methods were summarized narratively (Moons et al., 2025).

2.8 Data synthesis and statistical analysis

Meta-analysis was not performed because studies differed substantially in terms of populations, endpoints, input modalities, reference standards, algorithms, validation strategies, thresholds, and units of analysis. Findings were synthesized narratively and in structured tables. Performance estimates were reported as presented, with confidence intervals where available, and were not interpreted as evidence of superiority across studies with different endpoints, populations, or validation designs.

3 Results

3.1 Study selection

The search identified 2343 records (Scopus, 871; Embase, 626; Web of Science, 453; MEDLINE, 393). After removing 989 duplicates, 1354 records underwent title and/or abstract screening, of which 1309 were excluded at this stage. A total of 45 full-text reports were assessed, of which 35 were excluded for not developing an eligible ML model, leaving 10 studies for narrative synthesis (Fig. 1) (Yasin et al., 2025; Abbas et al., 2026; Álvaro-Afonso et al., 2025; Cakir et al., 2024; Cuce et al., 2023; Gao et al., 2025; Gudivaka et al., 2025; Kelly et al., 2024; Nie et al., 2026; Yোগapriya et al., 2022).

3.2 Overall study and model characteristics

The 10 included studies were published between 2022 and 2026 and used data from China (3), Türkiye (2), Spain, Australia (2), USA, and India. Most were retrospective or secondary analyses. Abbas et al. (2026) used radiographs collected prospectively for model development.

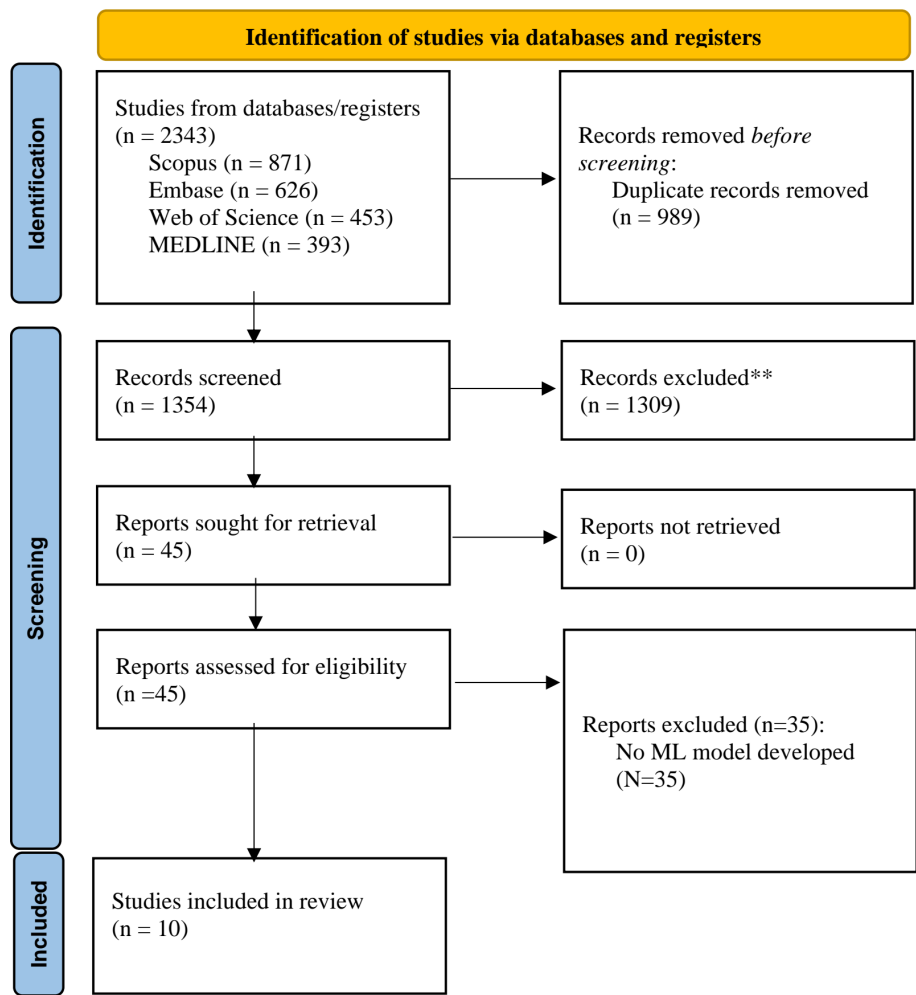


Figure 1. PRISMA flow diagram of study identification, screening, eligibility assessment, and inclusion.

Five studies directly addressed DFO diagnosis, while four evaluated complementary diabetic-foot endpoints using osteomyelitis as a predictor, severity component, or model context; also, one additional study evaluated model explainability. Among direct DFO studies, one used clinical/laboratory variables, two plain radiographs, and two MRI inputs. None evaluated incident DFO or DFO-specific treatment outcomes, including treatment response, healing, recurrence, treatment failure, amputation, or mortality. Study characteristics are summarized in Tables 1 and 2.

3.3 Models directly evaluating DFO

3.3.1 Binary DFO detection and classification

Three studies evaluated machine-learning models for binary DFO detection or classification (Abbas et al., 2026; Álvaro-Afonso et al., 2025; Yasin et al., 2025) Yasin et al. (2025) developed a two-center model using clinical and laboratory variables to differentiate DFO from soft-tissue diabetic-foot

infection (DFI) without bone involvement and reported both internal- and external-validation results (Yasin et al., 2025). Álvaro-Afonso et al. (2025) evaluated a ResNet-50 model using radiographs with ulcer-location information and histopathological confirmation from surgical bone biopsy (Álvaro-Afonso et al., 2025). Abbas et al. (2026) developed a deep-learning model for image-level DFO classification using plain radiographs with patient-level training and evaluation splits (Abbas et al., 2026) . Model inputs, reference standards, and performance measures are summarized in Tables 1 and 3.

3.3.2 Multiclass and differential diagnosis of DFO

Two studies evaluated MRI-based ML models for DFO classification and differential diagnosis (Cakir et al., 2024; Cuce et al., 2023). Cuce et al. developed a radiomics model using segmented T1- and T2-weighted MRI regions and a multilayer perceptron with cross-validation (Cuce et al., 2023). Cakir et al. (2024) evaluated deep-learning models based on

Table 1. Characteristics of studies with a direct diabetic-foot osteomyelitis endpoint. DFO: diabetic-foot osteomyelitis; DFI: diabetic-foot infection; CNO: Charcot neuropathic osteoarthropathy; CV: cross-validation; LR: logistic regression; MLP: multilayer perceptron; NR: not reported; ROI: region of interest; HbA1c: glycated hemoglobin; ESR: erythrocyte sedimentation rate; GLCM: gray-level co-occurrence matrix; GLDM: gray-level dependence matrix; SDHGLE: small-dependence high-gray-level emphasis; MRI: magnetic resonance imaging; GBM: gradient-boosting machine; CNN: convolutional neural network.

Study	Design/setting	Sample used in model and split	Model(s)	Endpoint	Reference standard	Variables/model inputs
Yasin et al. (2025); China	Retrospective, two-center diagnostic study (2015–2024).	$N = 3612$: DFO 1913; soft-tissue DFI without DFO 1699. Development $n = 3271$ (~ 2453 train/ ~ 818 internal validation); external $n = 341$.	LightGBM	DFO vs. soft-tissue DFI without bone involvement.	Composite clinical, laboratory, MRI, microbiology/histopathology, clinical course, and treatment response diagnosis.	Age, HbA1c, creatinine, albumin, sodium, ESR.
Cuce et al. (2023); Türkiye	Retrospective, single-center MRI radiomics study (2020–2022).	$N = 70$ patients: DFO 24; Charcot 17; trauma 29. T1 ROIs = 299; T2 ROIs = 301; 10-fold CV.	MLP (best); LR comparator	DFO one-vs.-all; secondary three-class DFO/Charcot/trauma differentiation.	DFO histology; Charcot clinical/laboratory follow-up and response to offloading; trauma controls.	14 selected T1/T2 radiomic features, including GLCM contrast, GLDM SDHGLE, percentiles, skewness, correlation, zone entropy, inverse variance, and sum average.
Cakir et al. (2024); Türkiye	Retrospective, single-center deep-learning study.	$N = 148$ patients: DFO 64; CNO 23; trauma 61. T1 ROIs = 679 (3395 augmented); T2 ROIs = 714 (3570 augmented); five repeated splits with 3-fold CV.	EfficientNet-b0 (best); ResNet-50 comparator	DFO one-vs.-all; secondary three-class DFO/CNO/trauma differentiation.	DFO histology; CNO clinical/laboratory response; trauma controls.	Radiologist-segmented T1/T2 MRI ROIs; learned CNN features (no hand-crafted predictors).
Álvaro-Afonso et al. (2025); Spain	Retrospective diagnostic study in a specialized diabetic-foot unit (2020–2024).	$N = 168$: DFO 123; no DFO 45. Reported 80 : 20 split (~ 134 train/ ~ 34 validation).	ResNet-50	Binary DFO detection on plain radiographs.	Surgical bone-biopsy histopathology.	Plain radiographs plus ulcer-location information; learned image features.
Abbas et al. (2026); Australia	Prospectively collected, single-center radiograph dataset; internal model development.	$N = 225$ patients/1227 images. Train 179 patients/981 images; validation 23/123; test 23/123. DFO-positive labels of images 628; negative 599.	DualBack-GFT	Primary image-level DFO classification; secondary lesion localization.	Consensus radiologist annotations and bounding boxes.	Dorsoplantar/oblique radiographs, expert bounding boxes, and learned global/local image features.

Table 2. Characteristics of complementary diabetic-foot ulcer-relevant studies. ABI: ankle-brachial index; DFU: diabetic-foot ulcer; DFO: diabetic-foot osteomyelitis; IWGDF/IDSA: International Working Group on the Diabetic Foot/Infectious Diseases Society of America; NR: not reported; XAI: explainable artificial intelligence; EHR: electronic health records; ML: machine learning.

Study	Design/setting	Sample used in model and split	Model	Endpoint	Reference standard	Variables/model inputs
Nie et al. (2026); China	Retrospective, two-center EHR study of hospitalized DFU patients (2020–2024)	$N = 800$. Development $n = 500$: train 350/test 150; independent external validation $n = 300$	Random forest	Concurrent wound-infection classification in DFU	IWGDF/IDSA clinical infection criteria; microbiology not mandatory	Age, diabetes duration, ABI, ulcer area, vascular complications, osteomyelitis
Gao et al. (2025); China	Retrospective, single-center hospitalized DFU cohort (2015–2024)	$N = 599$. Any-amputation model train/test 480/119; minor 440/109; major 354/88	XGBoost	Any, minor, and major in-hospital amputation	Index hospitalization amputation outcome	29 variables across demographics, comorbidities, inflammation, glycemia, renal/nutritional markers, wound severity, osteomyelitis, and vascular intervention
Yogapriya et al. (2022); India	Secondary analysis of an existing DFU photograph dataset	5890 original images augmented to 29 450 images; three train/test splits were evaluated (70 : 30, 80 : 20, and 90 : 10); the final training workflow used 21 204/5890/2356 images for training/validation/testing	DFINET CNN	Infected vs. non-infected DFU photographs	Senior-physician image label; exact clinical/microbiologic standard NR	Color DFU photographs only; no structured clinical variables
Gudivaka et al. (2025); USA	Computational DFU image classification study; dataset and setting unclear	Patient/image total and train/validation/test counts NR	Advanced ML practical method (unclear)	DFU severity/grade classification and infection clustering	Reference standard NR; osteomyelitis/bone-involvement definitions inconsistent	Mixed/unclear photographs, thermograms, or DICOM-derived images; no reliable final feature set
Kelly et al. (2024); Australia	Conference XAI/methods study using DFO radiograph patches	882 balanced patches (~ 441 DFO+/ ~ 441 DFO–); training/validation/test and patient-level counts NR	VGG16 + Mtree-CCEA	Counterfactual explanation quality for a fixed DFO classifier	Professionally labeled radiographs; clinical reference standard NR	224×224 12-bit radiograph patches; learned VGG16 features and counterfactual pixel changes

segmented MRI regions using transfer-learning architectures for DFO classification and differential diagnosis with Charcot neuroarthropathy and trauma (Cakir et al., 2024). MRI dataset characteristics, model inputs, validation approaches, and performance measures are summarized in Tables 1 and 3.

3.3.3 DFO localization and model explainability

In addition to image-level classification, Abbas et al. (2026) evaluated patch-based lesion localization and reported recall

of 82.4 %, precision of 42.2 %, and an F_1 score of 55.8 % (Abbas et al., 2026).

Kelly et al. (2024) evaluated an explainability approach on a previously developed deep-learning classifier for DFO detection from radiograph patches. The study concentrated on visual explanations and not diagnostic performance; although classifier accuracy was reported, AUC (area under the receiver operating characteristic curve), sensitivity, specificity, patient-level data separation, and external validation were not (Kelly et al., 2024).

Table 3. Model performance and validation for studies with a direct diabetic-foot osteomyelitis endpoint. DFO: diabetic-foot osteomyelitis; DFI: diabetic-foot infection; AUC: area under the receiver operating characteristic curve; CI: confidence interval; CV: cross-validation; NR: not reported; PPV/NPV: positive/negative predictive value; ROI: region of interest. “Yes” indicates that an internal-validation method was reported and does not by itself imply low risk of bias.

Study	Endpoint	AUC	Sensitivity	Specificity	Other reported performance	Internal validation	External validation
Yasin et al. (2025)	DFO vs. soft-tissue DFI without DFO	0.879 internal; 0.942 external	80.1 %	78.2 %	Accuracy 78.9 %; PPV 69.9 %; NPV 86.2 %	Yes – 75 : 25 holdout (2453/818) plus nested CV	Yes – independent center, $n = 341$
Cuce et al. (2023)	OM one-vs.-all, T2-weighted MRI, MLP	0.909	85.9 %	89.9 %	Accuracy 88.0 %; PPV 88.4 %; NPV 87.7 %	Yes – 10-fold CV on 299 T1 and 301 T2 ROIs; patient-level grouping NR	No
Cakir et al. (2024)	OM one-vs.-all, T1-weighted MRI, EfficientNet-b0	0.998	97.6 %	98.7 %	NR	Yes – five repeated splits with 3-fold CV on augmented ROIs; fixed counts NR	No
Álvaro-Afonso et al. (2025)	DFO present vs. absent	NR	92.8 %	4.4 %	Accuracy 69.1 %; PPV 72.6 %; NPV 18.2 %; F_1 score 81.4 %	Yes – reported 80 : 20 split (134/34)	No
Abbas et al. (2026)	Image-level DFO classification	0.968 (95 % CI 0.930–1.000)	90.5 %	95.0 %	Balanced accuracy 92.7 %; F_1 score 92.7 %	Yes – patient-level train/validation/test: 179/23/23 patients	No

3.4 Complementary DFO-relevant machine-learning studies

3.4.1 Osteomyelitis used as a model predictor

Nie et al. (2026) developed a random-forest model for wound infection classification in 800 hospitalized patients with DFU, with six final predictors including osteomyelitis. The AUCs were 0.853 in the internal test cohort and 0.832 in an independent external cohort (Nie et al., 2026).

Gao et al. (2025) developed XGBoost models for non-amputation, minor amputation, and major amputation outcomes among 599 hospitalized patients with DFU, with osteomyelitis included among 29 candidate variables. The AUCs were 0.772 for non-amputation, 0.800 for minor amputation, and 0.977 for major amputation (Gao et al., 2025).

3.4.2 Related infection and severity classification

Yogapriya et al. (2022) designed a convolutional neural network to classify infected versus non-infected DFUs using ulcer photographs. There was no separate assessment of osteomyelitis and no DFO-specific performance reported (Yogapriya et al., 2022).

Gudivaka et al. evaluated an ML framework for DFU severity classification, incorporating osteomyelitis or bone involvement into severity definitions. DFO was not assessed as an independent endpoint, and details on the dataset, reference standard, validation strategy, and DFO-specific performance were unavailable (Gudivaka et al., 2025).

3.5 Model inputs and predictor selection approaches

Direct DFO models utilized structured clinical and/or laboratory or imaging data; none combined clinical or laboratory variables with radiographic or MRI features. Reported predictors were demographic, glycemic, renal, inflammatory, and nutritional variables. Complementary diabetic-foot models included vascular, wound, inflammatory, and osteomyelitis-related variables. Predictors were selected by different methodologies across the studies, including expert

and statistical methods, tree-based importance, explainability, radiomic selection, and learned deep-learning features. Tables 1 and 2 summarize detailed predictors and modeling approaches.

3.6 Model validation and predictive performance

All five direct DFO studies reported internal validation using holdout; cross-validation; repeated data splitting; or separate training, validation, and testing groups. Only Yasin et al. (2025) performed independent external validation of a direct DFO model (Yasin et al., 2025). Among complementary studies, Nie et al. (2026) also externally validated its wound infection model in an independent hospital cohort (Nie et al., 2026).

AUCs among direct DFO models ranged from 0.879 to 0.998 but were not directly comparable because studies differed in terms of populations, endpoints, reference standards, modalities, case mix, units of analysis, and validation designs. Calibration was reported inconsistently, and several studies failed to report confidence intervals, clinically defined thresholds, or full sensitivity and specificity. Tables 3 and 4 summarize the performance and validation.

3.7 Risk of bias, quality, and applicability

The quality assessment is summarized in Table S1 in the Supplement. Most studies were judged at high overall risk of bias (applicability concerns were present in some studies), and methodological quality was variable. Inter-reviewer agreement was moderate to almost perfect for PROBAST + AI domains. Cohen's κ showed a substantial agreement for participants ($\kappa = 0.65$), a moderate agreement for predictors ($\kappa = 0.54$) and outcome ($\kappa = 0.55$), and an almost perfect agreement for overall applicability ($\kappa = 1.00$). Similar analysis and overall risk of bias ratings were seen across studies ($\kappa = 0.00$ due to limited rating variability; kappa paradox). All assessments were discussed, and a consensus was reached for use in the final analysis.

4 Discussion

In this systematic review of ML applications in DFO, we identified 10 studies evaluating models for direct DFO diagnosis, related diabetic-foot outcomes, or model interpretation. The available evidence was primarily focused on diagnostic applications, with direct DFO models using either structured clinical and laboratory variables or imaging-based inputs, including plain radiographs and MRI. Only one of the five direct DFO diagnostic studies reported independent external validation, and no included study developed a fully multimodal model integrating clinical or laboratory information with radiographic or MRI features. Overall, the current evidence does not establish a sufficiently consistent basis for comparable or generalizable ML approaches to DFO

given substantial heterogeneity in clinical endpoints, reference standards, input modalities, validation strategies, and reporting practices.

Although several included studies reported high discrimination, direct comparison of AUC, sensitivity, specificity, or other performance measures across models are not appropriate.

Among direct DFO diagnostic models, clinical and laboratory-based approaches demonstrated the potential value of routinely available patient-level information. These findings indicate that routinely collected clinical data can be useful for the classification of DFO, but the generalizability of such models relies on consistent definitions of outcomes, proper selection of predictors, and validation in different clinical settings. Prediction models may demonstrate strong discrimination when developed in specific populations but may lose performance when applied to patients with different case mix characteristics or clinical workflows (Collins et al., 2024a; Riley et al., 2024).

Interpretation of imaging-based models must consider the reference standard, the structure of the dataset, and the validation approach. Plain radiographs have limited sensitivity for early DFO and may show findings that overlap with other conditions affecting the diabetic foot. Therefore, careful selection of reference standards is essential when evaluating image-based models (Dinh et al., 2008; Senneville et al., 2008). Similarly, MRI-based models evaluated by Cuce et al. (2023) and Cakir et al. (2024) reported high internal discrimination for differentiating DFO from other bone-related conditions; however, these studies included relatively small cohorts and used segmented or augmented image data, limiting confidence in the generalizability of their performance estimates (Cakir et al., 2024; Cuce et al., 2023; Varoquaux and Cheplygina, 2022).

One major limitation was the lack of external validation. Cross-validation, holdout testing, and repeated data splitting evaluate performance on the available data but do not determine accuracy across institutions, populations, or clinical settings (Collins et al., 2024a). External validation is essential because changes in disease prevalence, patient characteristics, imaging protocols, laboratory measurements, and clinical decision-making pathways may significantly affect model performance (Riley et al., 2024). Furthermore, discrimination by itself does not establish clinical utility. Prediction models for clinical decision-making should also show reliable calibration, clinically meaningful thresholds, and evidence of clinical utility beyond statistical performance measures (Collins et al., 2024a; Efthimiou et al., 2024). In the present review, calibration reporting was inconsistent, and decision-analytic approaches were rarely evaluated, limiting the assessment of potential clinical impact.

The complementary diabetic-foot studies included in this review provided additional information regarding potential predictors and modeling approaches but should not be interpreted as DFO diagnostic models.

Table 4. Model performance and validation for complementary DFO-relevant studies. AUC: area under the receiver operating characteristic curve; CV: cross-validation; DFU: diabetic-foot ulcer; DFO: diabetic-foot osteomyelitis; MCC: Matthews correlation coefficient; NR: not reported. * For Gao et al. (2025), performance metrics are presented as reported by the authors for the any-, minor-, and major-amputation models, respectively.

Study	Endpoint	AUC	Sensitivity	Specificity	Other reported performance	Internal validation	External validation
Nie et al. (2026)	Wound infection in DFU	0.853 internal; 0.832 external	NR	NR	Internal CV AUC 0.820; external-cohort CV AUC 0.744	Yes – train/test 350/150 plus cross-validation	Yes – independent center, $n = 300$
Gao et al. (2025)	Non-amputation/minor amputation/major amputation*	0.772, 0.800, 0.977	75.7 %, 78.1 %, 98.7 %	77.8 %, 69.4 %, 88.9 %	NPV 66.0 %, 61.0 %, 88.9 %; PPV 84.8 %, 83.8 %, 98.7 %	Yes – 80 : 20 holdout plus 5-fold CV; test $n = 119/109/88$	No
Yogapriya et al. (2022)	Infected vs. non-infected DFU images	NR	90.57 %	93.49 %	Accuracy 91.98 %; precision 93.72 %; F_1 score 92.12 %; MCC 0.84	Yes – image-level 90 : 10 train/test split after augmentation; 27 094/2356	No
Gudivaka et al. (2025)	DFU grade/severity classification	NR	NR	NR	Maximum grade-specific classification accuracy 93.75 %	Unclear – validation procedure not reproducibly reported	No
Kelly et al. (2024)	Counterfactual explanation of DFO classifier	NR	NR	NR	Underlying classifier accuracy 78.5 %	Unclear – classifier split and patient partition	No

Current clinical assessment of DFO already relies on multiple complementary data sources, including wound characteristics, probe-to-bone findings, inflammatory markers, vascular status, plain radiography, and MRI when needed (Senniville et al., 2023). However, previous ML studies have generally evaluated these information sources separately: clinical and laboratory variables were used without imaging, while radiograph and MRI models were developed without broader patient-level information (Abbas et al., 2026; Álvaro-Afonso et al., 2025; Yasin et al., 2025; Cakir et al., 2024; Cuce et al., 2023). Future research should therefore investigate whether combining clinical and imaging data provides incremental value compared with unimodal approaches. Multimodal models should be developed and evaluated using the same patient population, standardized reference standards, and independent validation cohorts, allowing assessment of whether integration improves discrimination, calibration, and clinical utility rather than assuming benefit from increased model complexity alone (Collins et al., 2024a; Kline et al., 2022).

Future studies should focus on clinically available predictors, excluding variables collected after the intended prediction time or variables that directly contribute to the diagnostic reference standard and ensuring proper patient-level separa-

tion during imaging preprocessing and model development (Efthimiou et al., 2024; Varoquaux and Cheplygina, 2022). A clinically meaningful multimodal DFO model should be compared against clinical-only and imaging-only models, include external validation, report calibration and decision-analytic measures, and evaluate performance across relevant patient subgroups before implementation in clinical practice (Collins et al., 2024a, b; Riley et al., 2024).

This systematic review has several strengths. We performed a comprehensive search across multiple databases, used a structured extraction framework informed by CHARMS, separated direct DFO models from complementary diabetic-foot models, and synthesized evidence according to clinical endpoint rather than algorithm type. This approach reduced the risk of misusing general diabetic-foot-infection or severity models as DFO diagnostic models and offered a summary of model inputs, validation strategies, and reporting approaches. This review also has several limitations. There was considerable heterogeneity in populations, endpoints, reference standards, input modalities, and validation methods, which hampered quantitative synthesis and direct comparison of model performance. Furthermore, most studies had limited external validation and incomplete reporting of calibration, preprocessing procedures, and model de-

velopment details, limiting the assessment of reproducibility and clinical applicability.

5 Conclusions

Most of the ML research in DFO has focused on diagnostic classification using either structured clinical and laboratory variables or imaging-based approaches. However, the current evidence does not establish a sufficiently consistent or generalizable ML approach to DFO given substantial heterogeneity in study design, clinical endpoints, reference standards, input modalities, validation strategies, and reporting. None of the included studies developed a multimodal model that incorporated clinical or laboratory data with radiographic or MRI features, and none of the studies directly assessed prediction of incident DFO or DFO-specific treatment outcomes. Future research should establish standardized clinical endpoints and reference standards and focus on the development and validation of multimodal models integrating complementary information from clinical and imaging sources, testing clinical-only, imaging-only and combined approaches on the same cohorts, with external validation, calibration assessment, and clinical utility analysis prior to implementation in practice.

Data availability. All data used in this study were obtained from previously published articles cited in the paper. The extracted dataset supporting the findings of this study is available from the corresponding author upon reasonable request.

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/jbji-11-623-2026-supplement>.

Author contributions. FP: writing (original draft, review and editing), resources, methodology, investigation, data curation, conceptualization. SMAA: writing (original draft, review and editing), resources, methodology, investigation, data curation, conceptualization. FPe: writing (original draft, review and editing). TM: writing (original draft, review and editing). SAM: writing (review and editing), supervision. FB: writing (review and editing), supervision. MET and AJT: writing (review and editing), supervision. EFB: writing (review and editing), supervision, resources, project administration, methodology, investigation, conceptualization.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Journal of Bone and Joint Infection*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Ethical statement. Ethics approval was not required for this review, as the data do not provide personally identifiable information.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. Generative artificial intelligence tools were used solely for assistance with the grammar, punctuation, and typographical correction in initial versions of this paper. No content generation or scientific interpretation was performed using AI, and all final content was reviewed and approved by the authors.

Review statement. This paper was edited by Parham Sendi and reviewed by two anonymous referees.

References

- Abbas, S. Q., Saha, S., Dowling, J., Mehdizadeh, M., Vignarajan, J., Gupta, A., Frost, S., Nicolson, A., and Min, H.: Gated Backbone Fusion with Transformer Encoder for Diabetic Foot Osteomyelitis Screening and Localization in Radiographs, *J. Imag. Inform. Med.*, <https://doi.org/10.1007/s10278-026-01936-w>, 2026.
- Álvaro-Afonso, F. J., Tardáguila-García, A., López-Moral, M., Sanz-Corbalán, I., García-Morales, E., and Lázaro-Martínez, J. L.: Using Artificial Intelligence for Detecting Diabetic Foot Osteomyelitis: Validation of Deep Learning Model for Plain Radiograph Interpretation, *Appl. Sci.*, 15, 8583, <https://doi.org/10.3390/app15158583>, 2025.
- Armstrong, D. G., Tan, T.-W., Boulton, A. J., and Bus, S. A.: Diabetic foot ulcers: a review, *Jama*, 330, 62–75, 2023.
- Baker, J. C., Northrup, B. E., Ahlawat, S., Allen, H., Allison, G., Banks, J., Borloz, M. P., Chamarthy, M. R., Garner, H. W., Gross, C. E., Scott, J. A., Switaj, T., Zreloff, J., and Wessell, D. E.: ACR Appropriateness Criteria® Suspected Osteomyelitis of the Foot in Patients With Diabetes Mellitus: 2025 Update, *J. Am. Coll. Radiol.*, 22, S713–S737, <https://doi.org/10.1016/j.jacr.2025.08.031>, 2025.
- Cakir, M., Tulum, G., Cuce, F., Yilmaz, K. B., Aralasmak, A., Isik, M. İ., and Canbolat, H.: Differential diagnosis of diabetic foot osteomyelitis and charcot neuropathic osteoarthropathy with deep learning methods, *J. Imag. Inform. Med.*, 37, 2454–2465, 2024.
- Collins, G. S., Dhiman, P., Ma, J., Schluskel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G., and Van Smeden, M.: Evaluation of clinical prediction models (part 1): from development to external validation, *bmj*, e074819, <https://doi.org/10.1136/bmj-2023-074819>, 2024a.
- Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., and Van Smeden, M.: TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, *bmj*, e078378, <https://doi.org/10.1136/bmj-2023-078378>, 2024b.
- Cuce, F., Tulum, G., Yilmaz, K. B., Osman, O., and Aralasmak, A.: Radiomics method in the differential diagnosis of diabetic foot

- osteomyelitis and charcot neuroarthropathy, *Brit. J. Radiol.*, 96, 20220758, <https://doi.org/10.1259/bjr.20220758>, 2023.
- Dinh, M. T., Abad, C. L., and Safdar, N.: Diagnostic accuracy of the physical examination and imaging tests for osteomyelitis underlying diabetic foot ulcers: meta-analysis, *Clin. Infect. Dis.*, 47, 519–527, 2008.
- Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., and Salanti, G.: Developing clinical prediction models: a step-by-step guide, *bmj*, e078276, <https://doi.org/10.1136/bmj-2023-078276>, 2024.
- Gao, L., Liu, Z.-X., and Wang, J.-N.: Predictive model and risk analysis for outcomes in diabetic foot ulcer using eXtreme Gradient Boosting algorithm and SHapley Additive exPlanation, *World J. Diabet.*, 16, 104789, <https://doi.org/10.4239/wjd.v16.i7.104789>, 2025.
- Giurato, L., Meloni, M., Izzo, V., and Uccioli, L.: Osteomyelitis in diabetic foot: a comprehensive overview, *World J. Diabet.*, 8, 135–142, <https://doi.org/10.4239/wjd.v8.i4.135>, 2017.
- Guan, H., Wang, Y., Niu, P., Zhang, Y., Zhang, Y., Miao, R., Fang, X., Yin, R., Zhao, S., Liu, J., and Tian, J.: The role of machine learning in advancing diabetic foot: a review, *Front Endocrinol.*, 15, 1325434, <https://doi.org/10.3389/fendo.2024.1325434>, 2024.
- Gudivaka, R. K., Gudivaka, R. L., Gudivaka, B. R., Basani, D. K. R., Grandhi, S. H., and Khan, F.: Diabetic foot ulcer classification assessment employing an improved machine learning algorithm, *Technol. Health Care*, 33, 1645–1660, 2025.
- Harrison, H. S. G., Peters, E. J. G., Maas, M., Lavery, L. A., and Beekman, K. M.: Plain Radiography for Diagnosing and Monitoring Foot Osteomyelitis in Persons With Diabetes: Accuracy, Limitations and Clinical Utility, *Wound Repair Regen.*, 34, e70128, <https://doi.org/10.1111/wrr.70128>, 2026.
- Kelly, L., Masek, M., Lam, C. P., Abela, B., and Gupta, A.: Evolving Visual Counterfactual Medical Imagery Explanations with Cooperative Co-evolution using Dynamic Decomposition, in: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 291–294, <https://doi.org/10.1145/3638530.3654224>, 2024.
- Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., and Luo, Y.: Multimodal machine learning in precision health: A scoping review, *NPJ Digit. Med.*, 5, 171, <https://doi.org/10.1038/s41746-022-00712-8>, 2022.
- Llewellyn, A., Kraft, J., Holton, C., Harden, M., and Simmonds, M.: Imaging for detection of osteomyelitis in people with diabetic foot ulcers: A systematic review and meta-analysis, *Eur. J. Radiol.*, 131, 109215, <https://doi.org/10.1016/j.ejrad.2020.109215>, 2020.
- McDermott, K., Fang, M., Boulton, A. J., Selvin, E., and Hicks, C. W.: Etiology, epidemiology, and disparities in the burden of diabetic foot ulcers, *Diabet. Care*, 46, 209–221, 2023.
- Misir, A.: Artificial Intelligence and Machine Learning in Diabetic Foot Ulcer Care: Advances in Diagnosis, Treatment, Prognosis, and Novel Therapeutic Strategies, *J. Diabet. Sci. Technol.*, <https://doi.org/10.1177/19322968251363632>, 2025.
- Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., and Collins, G. S.: Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist, *PLoS Med.*, 11, e1001744, <https://doi.org/10.1371/journal.pmed.1001744>, 2014.
- Moons, K. G., Damen, J. A., Kaul, T., Hooft, L., Navarro, C. A., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., and Denaxas, S.: PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, *bmj*, e082505, <https://doi.org/10.1136/bmj-2024-082505>, 2025.
- Nie, X., Jiang, Y., Meng, X., Liu, J., Zhao, H., Chen, Y., Wang, J., and Lu, T.: Development and External Validation of a Machine Learning Model for Predicting Wound Infection in Diabetic Foot Ulcers, *Diabet. Metabol. Synd. Obes.*, 19, 586810, <https://doi.org/10.2147/DMSO.S586810>, 586810, 2026.
- Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., and Hagins, H.: Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021, *Lancet*, 402, 203–234, 2023.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., and Brennan, S. E.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *bmj*, n71, <https://doi.org/10.1136/bmj.n71>, 2021.
- Riley, R. D., Archer, L., Snell, K. I., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., and Collins, G. S.: Evaluation of clinical prediction models (part 2): how to undertake an external validation study, *bmj*, e074820, <https://doi.org/10.1136/bmj-2023-074820>, 2024.
- Senneville, E., Lombart, A., Beltrand, E., Valette, M., Legout, L., Cazaubiel, M., Yazdanpanah, Y., and Fontaine, P.: Outcome of diabetic foot osteomyelitis treated nonsurgically: a retrospective cohort study, *Diabet. Care*, 31, 637–642, 2008.
- Senneville, É., Albalawi, Z., van Asten, S. A., Abbas, Z. G., Allison, G., Aragón-Sánchez, J., Embil, J. M., Lavery, L. A., Alhasan, M., Oz, O., Uçkay, I., Urbančič-Rovan, V., Xu, Z.-R., and Peters, E. J. G.: IWGDF/IDSA Guidelines on the Diagnosis and Treatment of Diabetes-related Foot Infections (IWGDF/IDSA 2023), *Clin. Infect. Dis.*, ciad527, <https://doi.org/10.1093/cid/ciad527>, 2023.
- Snell, K. I., Levis, B., Damen, J. A., Dhiman, P., Debray, T. P., Hooft, L., Reitsma, J. B., Moons, K. G., Collins, G. S., and Riley, R. D.: Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA), *bmj*, e073538, <https://doi.org/10.1136/bmj-2022-073538>, 2023.
- Varoquaux, G. and Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future, *npj Digit. Med.*, 5, 48, <https://doi.org/10.1038/s41746-022-00592-y>, 2022.
- Yasin, P., Dong, S., Aizezi, Z., Yimit, Y., Yusufu, A., Yakufu, M., and Song, X.: 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, *BMC Med. Inform. Decis. Making*, 25, 420, <https://doi.org/10.1186/s12911-025-03236-9>, 2025.
- Yogapriya, J., Chandran, V., Sumithra, M., Elakkiya, B., Shamila Ebenezer, A., and Suresh Gnana Dhas, C.: Automated detection of infection in diabetic foot ulcer images using con-

volutional neural network, *J. Healthcare Eng.*, 2022, 2349849,
<https://doi.org/10.1155/2022/2349849>, 2022.