U.PORTO

MASTER'S DEGREE
PRIMARY HEALTHCARE

Systematic review of diabetic foot ulcer classification using artificial intelligence and machine learning techniques

Manuel Alberto Rodrigues Silva

M

2024

U.PORTO

FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO

Systematic review of diabetic foot ulcer classification using artificial intelligence and machine learning techniques

Author

Manuel Alberto Rodrigues Silva

Thesis Advisor

Sofia Baptista, MD, PhD

Thesis Co-advisor

Matilde Monteiro Soares, DPM, MSc, PhD

September of 2024

Investigators

Manuel Alberto Silva
USF Sanus Carandá, ULS Braga, Braga, Portugal
Faculty of Medicine, University of Porto, Porto, Portugal

Emma J Hamilton
Department of Endocrinology and Diabetes, Fiona Stanley Hospital, Murdoch, Australia
University of Western Australia, Medical School, Fiona Stanley Hospital, Murdoch, Australia

David A Russell
Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK
Leeds Vascular Institute, Leeds Teaching Hospitals NHS Trust, Leeds, UK

Fran Game
University Hospitals of Derby and Burton NHS Foundation Trust, Derby, UK

Sheila C Wang
Department of Medicine, Division of Dermatology, Women's College Hospital
University of Toronto

Sofia Baptista
MEDCIDS – Departamento de Medicina da Comunidade Informação e Decisão em Saúde; Faculty of Medicine, University of Porto, Porto, Portugal
CINTESIS@RISE - Center for Health Technology and Services Research; Faculty of Medicine, University of Porto, Porto, Portugal
CUF Porto Hospital, Portugal

Matilde Monteiro-Soares
Portuguese Red Cross Health School Lisbon, Lisbon, Portugal
Cross I&D, Lisbon, Portugal
MEDCIDS – Departamento de Medicina da Comunidade Informação e Decisão em Saúde; Faculty of Medicine, University of Porto, Porto, Portugal
CINTESIS@RISE - Center for Health Technology and Services Research; Faculty of Medicine, University of Porto, Porto, Portugal

Acknowledgments

To Prof. Sofia and Prof. Matilde – thank you for accepting the challenge to guide me through this journey. It was a privilege to have worked under your guidance. I could not have done it without your constant support and dedication.

To my family – thank for your love and care.

To Ana – thank you for always being there for me, even in the hardest times.

Table of contents

List of tables

List of figures

Abstract

Background: Diabetes-related foot ulceration (DFU) is a common complication of diabetes, with a significant impact on survival, healthcare costs and health-related quality of life. We aimed to systematically identify and collect the available evidence assessing the ability of machine learning (ML) based models in predicting clinical outcomes in people with DFU.

Methods: We searched PubMed, Scopus, Web of Science and IEEExplore for articles published up to July 2023. The literature was screened independently by two investigators for eligibility criteria and extracted data. Risk of bias was evaluated using Quality In Prognosis Studies (QUIPS) tool and Prediction Model Risk Of Bias Assessment Tool (PROBAST) by two investigators independently.

Results: A total of 11 studies, comprising 13 articles, were included in our systematic review, focusing on four outcomes: wound healing, lower extremity amputation (LEA), mortality and reintervention. Overall, 57 predictive models were created using mostly clinical characteristics, random forest as developing method and area under the curve (AUC) as discrimination accuracy measure. AUC varied from 0.56 to 0.94, with the majority of the models achieving a good discrimination ability (defined as AUC ≥ 0.8). All studies were found to have a high risk of bias, mainly due to lack of uniform variable definitions, outcome definitions and follow-up periods, insufficient sample sizes, and inadequate handling of missing data.

Conclusion: The ML-based models found in this study achieved good discrimination ability in people with DFU. However, models presented a high risk of bias meaning further studies with a stricter methodology are needed.

Key-words: Artificial intelligence, diabetic foot, classification, machine learning, prognosis

Resumo

Introdução: A úlcera podológica é uma complicação comum da diabetes, com impacto significativo na sobrevivência, nos custos de saúde e na qualidade de vida. O objetivo deste estudo foi identificar e recolher de forma sistemática a evidência disponível acerca das capacidades preditivas dos modelos baseados em *machine learning* (ML), no que toca a desfechos clínicos, em pessoas com úlceras de pé.

Métodos: Foi realizada uma pesquisa na PubMed, Scopus, Web of Science e IEEExplore por artigos publicados até julho de 2023. A literatura foi revista por dois investigadores de forma independente, tendo em conta os critérios de elegibilidade e a realização da extração de dados. O risco de viés foi avaliado através das ferramentas *Quality In Prognosis Studies* (QUIPS) e *Prediction Model Risk Of Bias Assessment Tool* (PROBAST) por dois investigadores de forma independente.

Resultados: Foram incluídos nesta revisão sistemática um total de 11 estudos, que compreenderam 13 artigos, onde foram abordados 4 desfechos clínicos: cicatrização, amputação do membro inferior, mortalidade e reintervenção. Ao todo, foram desenvolvidos 57 modelos preditivos utilizando maioritariamente características clínicas, o método *random forest* para desenvolvimento do modelo e a área sob a curva (AUC) como medida de discriminação. A AUC variou entre 0,56 e 0,94, com a maioria dos modelos a atingirem uma boa capacidade discriminativa (definida por AUC ≥ 0.8). Todos os estudos tinham um alto risco de viés, principalmente devido à falta de uniformidade na definição de variáveis preditivas, definições de desfechos e de períodos de seguimento, a amostras com tamanho insuficiente, e a uma gestão inadequada dos dados omissos.

Conclusão: Os modelos baseados em ML reportados neste estudo conseguiram atingir uma boa capacidade discriminativa em pessoas com diabetes e úlcera podológica. No entanto, os modelos apresentaram um alto risco de viés, indiciando que são necessários mais estudos com uma metodologia mais rigorosa.

Palavras-chave: Inteligência artificial, pé diabético, classificação, *machine learning*, prognóstico

Abbreviations

AMSTAR	A MeaSurement Tool to Assess systematic Reviews
AUC	Area Under the Curve
CI	Confidence Interval
CRP	C-Reactive Protein
DFU	Diabetes-related Foot Ulceration
EPV	Events Per Variable
IDSA/IWGDF	Infectious Diseases Society of America/International Working Group on the Diabetic Foot
IWGDF	International Working Group on the Diabetic Foot
LASSO	Least Absolute Shrinkage and Selection Operator
LEA	Lower-Extremity Amputation
Light GBM	Light Gradient Boosting Machine
LR	Logistic Regression
ML	Machine Learning
PAD	Peripheral Arterial Disease
PECO-S	Population, Exposure, Comparator, Outcome and Study type
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PROSPERO	International Prospective Register of Systematic Reviews
QUIPS	Quality In Prognosis Studies
RF	Random Forest
ROC	Receiver Operating Characteristic
SINBAD	Site, Ischemia, Neuropathy, Bacterial infection, Area and Depth
SVM	Support Vector Machine
WIfI	Wound, Ischemia, foot Infection

Preamble

In 2021, the estimated prevalence of diabetes between the ages of 20 and 79, in Portugal, was 14.1%, which corresponded to 1.1 million people. Lower extremity complications, which include diabetes-related foot ulcerations (DFU), were an important cause of morbidity and mortality, with 2,445 people undergoing amputation and 7.9% dying following amputation.[1]

Primary Care centres are the patients' first contact with the health system, and their role in preventing and treating chronic diseases, such as diabetes and its complications, is fundamental. In 2022, 8.4% of the registered patients in Portuguese Primary Care were people with diabetes.[2]

An annual diabetic foot assessment and optimal management by a multidisciplinary team, including family physicians, nurses and podiatrists, can reduce the need for hospitalization, mortality and amputation.[3] In fact, a systematic review found that 94% of the studies reported a reduction in major amputations after the institution of a multidisciplinary team (i.e., two or more types of clinicians working together).[4]

According to Portuguese Primary Care data, in 2022, an annual foot assessment was made in 74% of patients; of these, 10% were classified as having moderate to high risk of DFU. In total, 0.25% of patients were diagnosed with a DFU.[2]

In the subject of diabetes-related foot complications, the International Working Group on the Diabetic Foot (IWGDF) is a worldwide recognised group that has been producing evidence-based guidelines to inform healthcare providers on strategies for the prevention and management of diabetic foot disease. In 2023, all IWGDF guidelines were updated based on systematic reviews of the literature and the formulation of recommendations by multidisciplinary experts. One of these guidelines concerned the classification of DFUs.[5] However, the corresponding systematic review focused only on classifications developed using more "classical" methods.

To update the recommendations for the classification of DFUs in the future, the IWGDF team sought to conduct a systematic review focusing on the predictive abilities of models based on machine learning techniques. Due to the inability to conduct it until the 2024 publications, this review was put on hold.

The Chair and Secretary of this group have invited me to update the search and develop the systematic review. As a master's student in Primary Healthcare and a Medical Resident in Family Medicine, I was privileged to participate in this study.

Primary care must manage many interconnected and interdependent issues, which are, by definition, complex. Some believe that artificial intelligence and machine learning may help manage such complexity.[6] Therefore, this study allowed me to learn more about machine learning algorithms and their ability to predict individual patients' outcomes based on variables commonly used in clinical practice. It is my conviction that these tools play an important part in changing how Primary Care providers approach patients with DFU.

Introduction

Diabetes is a rapidly growing disease. Since 2000, the prevalence of diabetes has more than tripled, reaching, in 2021, 10.5% of the adult population in the world.[7]

The increase in diabetes prevalence is associated with the rise of its related complications.[8] Diabetes-related foot ulceration (DFU), defined as a break in the skin of the foot that involves at least the epidermis and part of the dermis[9], is the most commonly recognised complication affecting the lower extremities. The risk of a person with diabetes developing a DFU across their lifetime is around 19 to 34%.[10] Approximately 20% of people who develop a DFU will require lower-extremity amputation (LEA)[10], and 10% will die within one year of their first DFU diagnosis.[11,12]

In the United States, foot complications contribute to $273 billion in direct costs and $90 billion in indirect costs.[13] Apart from the impact of a DFU on mortality and healthcare costs, people with DFUs also have a significantly lower health-related quality of life.[14]

The evaluation and prognosis of a DFU varies considerably according to person, limb and ulcer-related characteristics. For that reason, classification and scoring systems were developed to create groups of patients with similar characteristics to whom similar levels of care would apply. Furthermore, they can be used to communicate wound and person characteristics between professionals, estimate an individual's prognosis, help in clinical practice decision-making, and audit and compare populations.

A systematic review from the International Working Group on the Diabetic Foot (IWGDF) in 2023 found 28 different classification and scoring systems for DFUs.[15] No gold standard exists, despite the many classification and scoring systems available. Therefore, each system should be used according to the intended purpose, available resources, expertise, and clinical setting. In the IWGDF 2023 updated guidelines[5] several recommendations were made: SINBAD system (Site, Ischemia, Neuropathy, Bacterial infection, Area and Depth) was recommended for communication between healthcare professionals and clinical audits; IDSA/IWGDF (Infectious Diseases Society of America/International Working Group on the Diabetic Foot) was recommended for infected ulcers; and WIfI (Wound, Ischemia, foot Infection) was recommended for peripheral arterial disease (PAD) or for when resources and expertise are available. No available classification and scoring system for individual prognostication was recommended.

Expert opinion and conventional statistical methods, such as linear regression and other generalised linear models, have been used to develop classifications to help predicting in people with DFU their clinical outcome.[16,17] However, these methods lack in detail and do not capture the complex non-linear relationships between risk factors and outcomes, compromising the classification systems' predictive ability.

Recent technological advances have allowed the use of machine learning (ML) strategies, a branch of artificial intelligence, in healthcare. ML algorithms can use data from several sources, capture complex patterns and thus may perform better than traditional models[18,19], – especially in settings with high variability.

ML algorithms can be divided into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.[20,21] In supervised learning, the system infers a function from labelled training data and a collection of training examples. The most common are classification models (predict classes) and regression models (predict numerical values).[20,21] In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables without the need for human interference, commonly using clustering and association.[20,21] In semi-supervised learning, there is a combination of supervised and unsupervised methods, and the system works with labelled and unlabelled data.[20,21] In reinforcement learning, the system tries to learn through direct interaction with the environment.[20,21]

ML has been applied successfully to healthcare. A systematic review from Kavakiotis and colleagues searched the applications of ML in diabetes research and found that most of the algorithms (85%) used supervised approaches, usually when performing prediction tasks.[22] When it comes to DFU care, a systematic review by Tulloch and colleagues found multiple applications of ML, namely in classification, image analysis and segmentation.[23] However, this review focused on identifying the presence and type of DFU but not predicting clinical outcomes in people with DFUs.

Our systematic review aimed to collect all the available evidence assessing the prognostic abilities of ML-based models in predicting clinical outcomes in people with DFUs. We focused on the comparison between models, their performance and discussed their applicability in the DFU care context. We hope they can facilitate decision-making and debate the importance of integrating this type of model into daily clinical practice worldwide.

Methods

This systematic review was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis)[24] guidelines, and we used the AMSTAR (A MeaSurement Tool to Assess systematic Reviews)[25] tool to verify if the most important aspects have been included in our systematic review. We registered our review in the PROSPERO (International Prospective Register of Systematic Reviews) database in July 2022 and updated in August 2023 under CRD42022308248.

Search strategy

We searched the MEDLINE database (PubMed), Scopus, Web of Science and IEEExplore in two phases. In the first phase, we performed a search on 26 February 2022 to identify all studies published with no beginning date until December 2021 (inclusive). In the second phase, we updated the search on 10 August 2023 to identify all studies published from January 2022 until July 2023 (inclusive). No restrictions were applied.

Search queries are available in Supplementary Table 1. To refine our query, we have used as "satellite" some pertinent articles included in the systematic review from Tulloch and colleagues that addressed a similar topic.[23]

Reference lists of the included articles and previous systematic reviews were reviewed to find additional pertinent articles. Experts in the area were contacted to identify any other articles not identified by our query, namely internal medicine physicians, vascular surgeons, endocrinologists, nurses, podiatrists, and human movement scientists.

Inclusion and exclusion criteria

Studies were selected based on the PECO-S elements (population, exposure, comparator, outcome and study type), as explained below.

Population

Articles were considered eligible if the population included at least 80% of adults with diabetes and a foot ulcer or if a subgroup analysis of such participants was provided.

Exposure/comparator

We defined the exposure of interest as being classified at higher risk (exposure) or lower risk (comparator) by any model developed using artificial intelligence techniques to predict outcomes by assessing more than one patient, foot and/or ulcer characteristic.

We also investigated the association between the models' composing variables (i.e., each variable included in the model) and the different outcomes.

Outcomes

The authors selected the outcomes for the study using the list provided in the systematic review by Dovell and colleagues as a foundation[26]. Definitions of the outcomes were made according to the document from the IWGDF.[9]

Our primary outcome was wound healing: reaching intact skin, meaning complete epithelialisation without any drainage of a previous foot ulcer site.

As secondary outcomes, we used the following:

a) lower-extremity amputation (LEA): resection of a segment of a limb through a bone or a joint;
b) hospitalisation: care in a hospital that requires admission as an inpatient and usually requires an overnight stay;
c) length of stay: period of time in which a person is committed to a hospital;
d) health-related quality of life: a person's perceived physical and mental health;
e) survival: the state or fact of continuing to live or exist;
f) ulcer-free survival period/time: period of time in which a person is alive and without a foot ulcer;
g) LEA-free period: period of time in which a person is alive and without a LEA.

Study type

We included analytic longitudinal anterograde studies, meaning clinical trials and cohort studies. If a study was presented as an abstract or poster, further searching was done to identify if it gave origin to a full article. If not, the study was excluded.

Eligibility assessment and data extraction

Articles were included if they fulfilled all the eligibility criteria mentioned above.

The search was conducted in two phases. In both phases, the studies were reviewed independently by two reviewers: Matilde Monteiro-Soares (MMS) and David Russell (DR) or Emma Hamilton (EH) in the first phase, MMS and Manuel Alberto Silva (MAS) in the second phase. Studies were selected based on their titles and abstracts in the first stage and the complete text of the articles in the second stage. Divergent opinions were resolved by consensus. We used EndNote 20® to manage references and identify duplicates. Subsequently, we used Rayyan QCRI[27] for the blind and independent selection of references to be included in our systematic review. The proportion of agreement between the two reviewers was calculated for each stage.

Data was extracted from each included study using a spreadsheet and summarised in tables that included the following information: (i) article identification (namely authors, year of publication, country where study was conducted), (ii) methods (namely study design, inclusion of participants, sample size, follow-up, context of study), (iii) model characteristics (namely purpose, methods for development, validation

conducted, variable definitions), (iv) outcome definition, and (v) results and analysis (namely participants' age, type of diabetes, diabetes duration, gender, measures and statistical methods used).
Data was extracted by one reviewer (MAS) and confirmed by a second reviewer (MMS). Divergent opinions were solved by consensus.

Risk of bias

The risk of bias was assessed using the Cochrane Risk-of-Bias (RoB 2) tool[28] for randomised trials for impact analysis. If a study had a low risk of bias in all five domains, it was classified as at low risk of bias; if some concerns exist in at least one domain without any domain with a high risk of bias, it was classified as with some concerns; and if at least one domain has high risk of bias or some concerns exist for multiple domains, it was classified as at high risk of bias.
For observational longitudinal studies of clinical prognosis, both the Quality In Prognosis Studies (QUIPS) tool[29] and the Prediction model Risk Of Bias ASsessment Tool (PROBAST)[30] were used. In the case of QUIPS, we evaluated the studies according to five of the six proposed domains. We considered that the study confounding domain was not pertinent as the present paper aims to study the association between variables and outcomes regardless of a causal relationship. Thus, this domain was classified as low risk in all studies. Overall, if a study had a low risk of bias in the six domains, it was classified as being at very low risk of bias, in four to five domains as low risk of bias, and in three or fewer domains as high risk of bias. In the case of PROBAST, if a prediction model had a low risk on all domains relating to bias and applicability, it was classified as at low risk of bias or low concern regarding applicability; if a model had a high risk for at least one domain, it should be classified as having high risk of bias or high concern regarding applicability; if a model had unclear risk in one or more domains and had low risk in the remaining domains, it may be classified as having unclear risk of bias or unclear concern regarding applicability.
Two reviewers (MMS and MAS) assessed the risk of bias. Divergent opinions were resolved by consensus. The proportion of agreement between the two reviewers was calculated.

Results

Search results

In the first phase, we retrieved a total of 1950 references after removing the duplicates. A total of 90 references were selected in the first stage (title and abstract screening), with a proportion of agreement of 95% among the assessors (MMS, DR and EH). After the second stage (full-text screening), with an agreement of 98%, four references were included in our systematic review.

In the second phase (update of search), we retrieved 462 additional references. A total of 81 references were selected in the first stage, with a proportion of agreement of 90% between the two assessors (MMS and MAS). After the second stage, we included seven additional references in our systematic review, with an agreement of 98%. From searching the references of previous reviews (systematic or not), of included studies, and from contacting experts, we retrieved an additional two references. Thus, we included 13 references corresponding to 11 studies (see Figure 1).

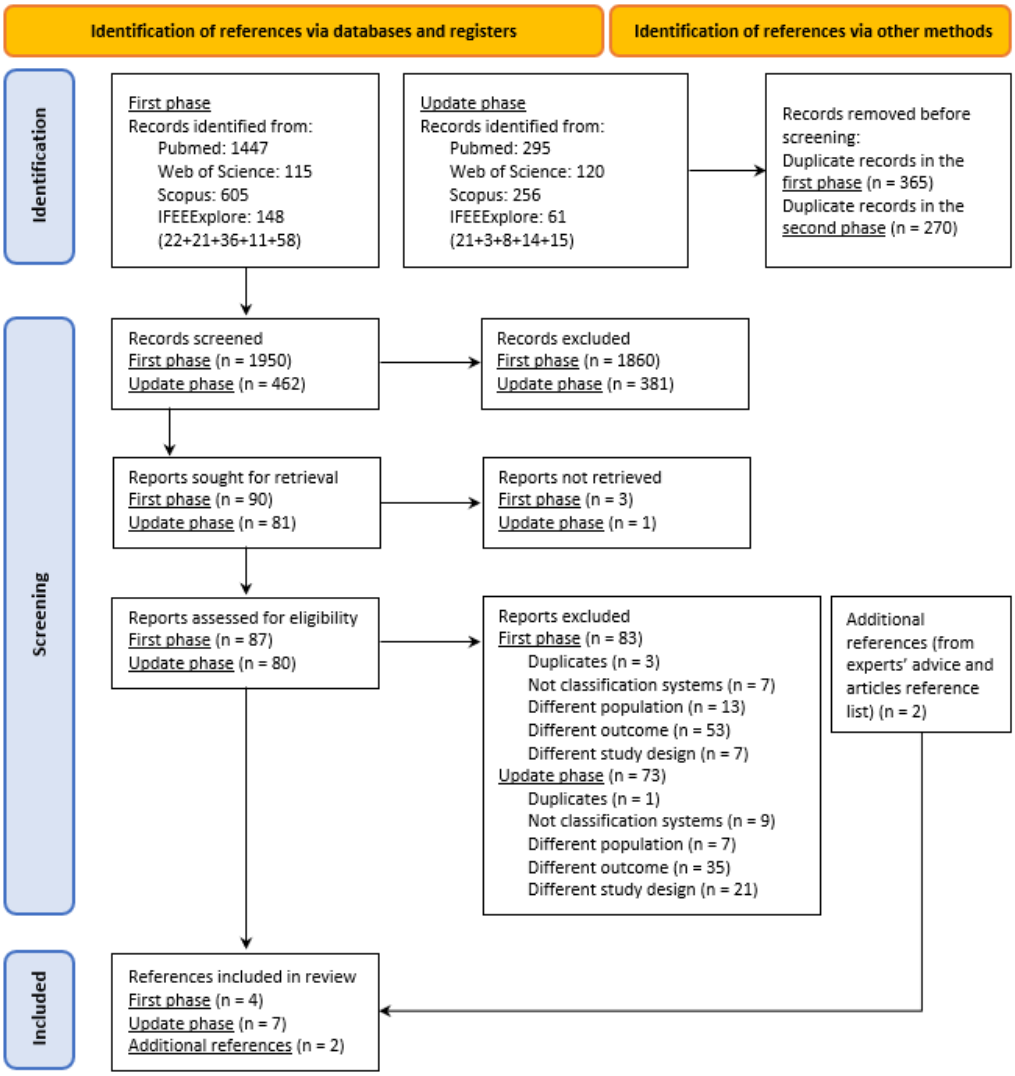

Figure 1. Articles' selection process.

Studies' design, setting and population

We have included studies published between 2016 and 2023, conducted in five countries: the United States (n=5, 45%),[31-35] China (n=3, 27%),[36-39] Germany (n=1, 9%),[40,41] Poland (n=1, 9%),[42] and India (n=1, 9%).[43] Eight studies (73%) were retrospective cohorts,[31,33-39,43] and three were prospective cohorts.[32,40-42] Six studies (55%) were single-centre,[34,36,39-43] with sample sizes ranging from 46 to 618 participants (median of 201), while five studies were multicentre,[31-33,35,37,38] and had a sample size varying from 204 to 88,898 participants (median of 53,354).

We have separated the results by the clinical outcome (wound healing, LEA, mortality and reintervention) and organised the studies included in each table (Table 1 to 4) by a higher stage of development (meaning external validation, internal validation or derivation only), study design with less risk of bias (meaning prospective cohort, or retrospective cohort study), multicentre versus single centre, and larger sample size.

Prediction models' characteristics by clinical outcome

Wound healing

A total of five papers (38%)[32,34,35,37,42] used wound healing as an outcome with some variations (see Table 1): two studies[32,34] evaluated wound healing; one study assessed delayed wound healing[35]; one study evaluated hard-to-heal wound[37]; and one study[42] evaluated wound healing failure. Follow-up periods for the mentioned outcomes ranged from four to 16 weeks, while one study[34] did not explicitly define any period of time for measuring the outcome.

Four studies[32,34,37,42] included only DFUs, with a total of 846 participants. Jung et al.[35] included several types of wounds, with 6,055 (4% of the overall sample) being neuropathic DFUs and provided a subgroup analysis. The incidence of wound healing varied from 35.1 to 78.8%. Regarding hard-to-heal wounds and wounds with delayed or failed healing, the incidence ranged from 11.6 to 66.0%.

A median of 35 clinical variables per study were assessed for model construction. Final models included between four and 865, with a median of 10 variables per study, distributed across four categories: demographic characteristics, medical history, laboratory data and foot-related characteristics (see Supplementary Table 2). Kim et al.[34] also included image-based characteristics retrieved from photographs through the user's subjective observation and deep learning techniques. The most commonly included were wound area (n=4, 80%), gender (n=3, 60%) and C-reactive protein (CRP) (n=3, 60%).

Regarding ML methods, four studies (80%)[34,35,37,42] employed multiple techniques simultaneously. The most applied ML method was random forest (RF), which was used in four studies (80%), followed by support vector machine (SVM) and least absolute shrinkage and selection operator (LASSO) regression, which was used in two studies each (40%).

Overall, across the five papers, 20 prediction models were created. Missing data was excluded in two studies[35,37], and was not reported in one study.[42] One study[32] reported the existence of missing values but not the approach to handle them, and one study[34] used imputation with a k-NN algorithm. As for model validation, every study used internal validation processes. Calibration was evaluated in two studies[32,35] using the Hosmer-Lemeshow test and Brier score, and discrimination accuracy was assessed by several measures including the area under the curve (AUC). Reported AUCs in model testing ranged from 0.636 to 0.864. The model which showed apparent better discrimination was developed in the study by Wang and colleagues,[37] using 10 clinical variables and the naïve Bayesian classifier. However, 95% confidence intervals (CI) were never reported.

Table 1. Characteristics of the included studies organized by model development stage, study design, setting and sample size: wound healing as outcome.

Reference	Study design, setting and follow-up	Study population and characteristics	Development/ validation	Variables assessed	Primary and secondary outcomes	Results	Comments
Margolis, 2022[32]	Prospective cohort Multicenter 4 wound care centers in USA Inclusion period: NR Follow-up period: 16 weeks	n = 204 people with plantar DFU n=1 (0.5%) lost to follow-up n=2 (1.0%) insufficient follow up Mean age: 58 years 73% male Mean diabetes duration: 18 years Inclusion criteria: adult-onset DM; DFU on the plantar aspect of the foot eligible for standard care; adequate arterial flow for healing; ≥ 40 years at the time of DFU diagnosis	Development (train): LASSO regression model Calibration: Hosmer-Lemeshow statistics Internal validation (test): 10-fold cross-validation	26 variables assessed; 4 variables included in the model wound duration, wound area, BMI, adequate arterial flow	Wound healing at 16 weeks	35.1% healed AUC: 0.7212	Inclusion period NR Missing values reported, but approach NR Information provided does not allow model application 95% CI not reported No external validation
Poradzka, 2023[42]	Prospective cohort Single center University hospital in Poland Inclusion period: February 2018 to March 2019 Follow-up period: 3 months	n = 164 consecutive people admitted with DFU (with 199 feet with DFU) Mean age: 61 years 75% male Mean diabetes duration: 17 years Inclusion criteria: people admitted with at least one DFU (according to IWGDF definitions) Exclusion of n=13 (7.3%) 2 did not agree to take part in study 2 withdrew consent 2 lost to follow-up 7 died during the follow-up	Development (train): Artificial neural network (70% for teaching and 15% for validating of sample randomly selected) Logistic regression Internal validation (test): 15% of sample randomly selected	35 variables tested in univariate or multivariate analysis, reducing the group to 6 (based on significance): probe-to-bone test, ESR, CRP, foot skin temperature, DFU area, DFU duration Added to the model (based on previous study): prior amputation, presence of blood flow in the Doppler probe, ABI	Healing failure (versus complete healing with or without LEA)	66% of ulcers without complete healing ANN AUC: 0.85 Accuracy: 82.9% Sensitivity: 91.6% Specificity: 66.2% PPV: 84.5% NPV: 80.4% F1 score: 87.9% Logistic regression Teaching AUC 0.73 Testing AUC 0.67	Single center study Analysis by limb and not person n=4 healed after minor LEA and were included in the healed success group No missing values reported Excluded lost to follow-up Logistic regression model not provided No external validation 95% CI not reported No calibration measures reported

Jung, 2016[35]	Retrospective cohort Multicenter 68 Healogics wound care centers in USA Inclusion period: 2009 to 2013 Follow-up period: mean 52 days	n= 6055 neuropathic DFUs (4% of overall sample) Mean age: 66 years 53% male Mean diabetes duration: NR	Development (train): 80% of overall sample LASSO regression; random forest; gradient boosted tree Calibration: Brier reliability of 0.00018 Internal validation (test) 20% of overall sample	865 predictors which sum up in 52 main variables: centre code, gender, age, palliative care, complex care, area*, length*, width*, depth, grade, thickness, temperature, exudate, margin, epithelization, tunnelling, undermining, exposed bone, exposed joint, exposed muscle, exposed tendon, brawny induration, oedema, excoriation, callus, crepitus, fluctuance, friable, rash, atrophy blanche, cyanosis, ecchymosis, erythema, hemosiderosis, pallor, rubor, tenderness on palpation, slough, delayed recurrence, result of accident, wound recurrence, clustered wound, pending amputation on presentation, granulation quality, granulation amount, primary insurance, number of wounds, duration of care, previous wound count, ICD9 codes, wound type, wound location (all variables were collected at onset of care; variables with asterisks were collected at onset of care and 1 week later)	Delayed wound healing (≥15 weeks to closure)	11.6% delayed wound healing Best was gradient boosted tree (only one reported) AUC: 0.823	Retrospective Follow-up period not clear Patients lost to follow-up were excluded Only 14% of wounds DFUs and only for neuropathic DFUs (4%) exists subgroup analysis without stratification in random split sample Not clear of number of people with neuropathic DFU Calibration measures only for overall sample Outliers and missing values were excluded No external validation 95% CI not reported

Wang, 2022a[37]	Retrospective cohort Multicenter 2 tertiary hospitals in eastern China Inclusion period: January 2018 to December 2019 Follow-up period: NR	n = 362 people with DFU > 63 years 71% male >12 years duration Inclusion criteria: 18–89 years, University of Texas Grade 3 DFU Exclusion criteria: tumour-induced ulcers; previous major LEA; abandoned treatment; had incomplete information	Development (train): 70% of sample randomly selected 6 different models General linear regression Naïve Bayesian Support vector machine Random forest k-nearest neighbour Adaptive boosting Internal validation (test): 30% of sample randomly selected	21 variables assessed 10 variables selected: 6 selected after univariate regression analysis (sex, random blood glucose, diabetic retinopathy, peripheral arterial disease, smoking history, and CRP); 4 selected based on literature and experience (insulin use, wound area, serum albumin, and serum creatinine)	Hard-to-heal DFU (PAR < 50% by week 4)	54% hard-to-heal DFU Adaptive boosting Accuracy: 0.713 PPV: 0.63 Sensitivity: 0.674 F1 score: 0.667 AUC: 0.804 General linear regression Accuracy: 0.750 PPV: 0.654 Sensitivity: 0.791 F1 score: 0.68 AUC: 0.826 K-nearest neighbor Accuracy: 0.713 PPV: 0.611 Sensitivity: 0.767 F1 score: 0.688 AUC: 0.817 Naïve Bayesian (model https://predicthardtoheal.azurewebsites.net) Accuracy: 0.750 PPV: 0.629 Sensitivity: 0.907 F1 score: 0.744 AUC: 0.864 Random forest Accuracy: 0.722 PPV: 0.614 Sensitivity: 0.814 F1 score: 0.66 AUC: 0.827 Support vector machine Accuracy: 0.759 PPV: 0.660 Sensitivity: 0.814 F1 score: 0.694 AUC: 0.819	Retrospective Those lost to follow-up and with missing values were excluded 95% CI not reported No calibration measures reported No external validation

Kim, 2020[34]	Retrospective cohort Single-center Podiatry and wound clinic USA Inclusion period: Nov 2014 to Jul 2017 Follow-up period: NR	n = 113 people with DFU (208 DFU) Mean age: >58.5 years % male not clear Mean diabetes duration: NR Exclusion criteria: previously infected DFU resulting in complete epithelization; previously DFU marked as healed that re-ulcerated at same location; DFU without image at initial visit 42 people (173 DFU) excluded	Development (train): 75% of sample random forest and support vector machine Internal validation (test): 25% of sample 3-fold cross-validation and grid-search	2133 Image features (DFU manually segmented from photographs): 85 hand-crafted image features (colour and texture) plus 2048 deep learning-based features (extracted from the GAP layer of ResNet50) 48 clinical variables: wound length, width, depth; foot, age, sex, ethnicity, CCI, diabetic retinopathy, lymphocyte count, HbA1c, albumin, pre-albumin, CKD stage coded, CRP, ESR, infection, probe-to-bone test, X-ray, MRI performed, Tc99 performed, dorsalis pedis pulse measured, posterior tibial pulse measured, ankle systolic pressure measured, toe systolic pressure measured, $TcPO_2$, BSA, BMI, TCC, offload, immunosuppressants, oral steroids, antihypertensives, oral hypoglycemics, canagliflozin, insulin, heparin, allopurinol, NSAIDs, aspirin, warfarin, Xa inhibitors, race, University of Texas stage, grade	Healing (healed or not healed)	78.8% of DFU healed Models built with all available features Development Random forest AUC: 0.734 Accuracy: 0.811 PPV: 0.828 Sensitivity: 0.923 F1: 0.873 Support vector machine AUC: 0.734 Accuracy: 0.811 PPV: 0.828 Sensitivity: 0.923 F1: 0.873 Validation Random forest AUC: 0.691 Accuracy: 0.671 PPV: 0.743 Sensitivity: 0.823 F1: 0.777 Support vector machine AUC: 0.735 Accuracy: 0.726 PPV: 0.780 Sensitivity: 0.854 F1: 0.814 Models built only with image features Development Random forest AUC: 0.760 Accuracy: 0.811 PPV: 0.852 Sensitivity: 0.885 F1: 0.868 Support vector machine AUC: 0.794 Accuracy: 0.784 PPV: 0.909 Sensitivity: 0.769 F1: 0.833 Validation Random forest	Retrospective Single-center study Patient follow-up period NR High rate of patient exclusion Those lost to follow-up were excluded Not described how sample was split for train and test 95% CI not reported Correspondence between people and DFU not explained or if analysis by DFU No calibration measures reported No external validation

			AUC: 0.683 Accuracy: 0.646 PPV: 0.795 Sensitivity: 0.671 F1: 0.727 Support vector machine AUC: 0.691 Accuracy: 0.561 PPV: 0.840 Sensitivity: 0.471 F1: 0.600 Models built only with clinical features Development Random forest AUC: 0.636 Accuracy: 0.784 PPV: 0.765 Sensitivity: 1.000 F1: 0.867 Support vector machine AUC: 0.657 Accuracy: 0.703 PPV: 0.800 Sensitivity: 0.769 F1: 0.784 Validation Random forest AUC: 0.693 Accuracy: 0.707 PPV: 0.757 Sensitivity: 0.875 F1: 0.805 Support vector machine AUC: 0.701 Accuracy: 0.591 PPV: 0.810 Sensitivity: 0.567 F1: 0.652 Models built only with deep learning features Development Random forest AUC: 0.670 Accuracy: 0.757 PPV: 0.793 Sensitivity: 0.885

			F1: 0.836 Support vector machine AUC: 0.670 Accuracy: 0.757 PPV: 0.793 Sensitivity: 0.885 F1: 0.836 Validation Random forest AUC: 0.692 Accuracy: 0.720 PPV: 0.764 Sensitivity: 0.880 F1: 0.815 Support vector machine AUC: 0.726 Accuracy: 0.707 PPV: 0.737 Sensitivity: 0.912 F1: 0.813 Random forest + support vector machine models with imaging features alone outperformed models trained with deep learning-based features alone (p = 0.013) Support vector machine models trained with imaging features alone outperformed models with clinical features alone (p = 0.031)

ABI: ankle-brachial index; ANN: artificial neural network; positive predictive value; AUC: area under the curve; BMI: body mass index; BSA: body surface area; CCI: Charlson comorbidity index; CI: confidence interval; CKD: chronic kidney disease; CRP: C-reactive protein; DFU: diabetic foot ulcer; DM: diabetes mellitus; ESR: erythrocyte sedimentation rate; GAP: global average pooling; ICD-9: International Classification of Diseases, 9th Revision; IWGDF: International Working Group on the Diabetic Foot; LASSO: least absolute shrinkage and selection operator; LEA: lower extremity amputation; MRI: magnetic resonance imaging; NPV: negative predictive value; NR: not reported; NSAIDs: non-steroidal anti-inflammatory drugs; PAR: percent area reduction; ResNet50: residual network 50; Tc99: technetium-99 bone scan; TCC: total contact cast use; $TCPO_2$: transcutaneous oxygen pressure; USA: United States of America.

Lower extremity amputation

Seven articles (54%)[31,33,36,38-41,43] had LEA as outcome (see Table 2). One study (14%) focused on major LEA[31], another study (14%) on minor LEA[38], and four studies (57%) assessed simultaneously two different types of LEA (minor and major, and major and any).[33,36,40,41] The remaining papers focused on any form of amputation. The follow-up period for determining LEA occurrence ranged from six to 12 months; four studies[31,36,38,39] did not mention a predefined follow-up period for determining the outcome.

A total of 417,315 people with diabetes were included. The incidence of major LEA ranged from 5.9 to 12.2%, whereas of minor LEA ranged from 11.5 to 20.7%. Concerning any form of LEA, the incidence varied from 1.6 to 31.6%.

A median of 21 clinical variables per study were assessed for model construction. Final models included between seven and 37, with a median of 10 variables per study. Du et al.[39] only reported the most relevant variables to model construction, so only those were accounted for. The most frequent variables were age (n=5, 71%), and gender, diabetes duration, smoking history, haemoglobin A1c, creatinine, albumin and random blood glucose (all n=3, 43%) (see Supplementary Table 3).

Regarding ML techniques, four studies (57%) employed multiple techniques simultaneously,[31,33,38,39]. The most used ML method was RF, which was used in four studies (57%).[31,33,38,39]

Overall, across the seven papers, 27 prediction models were created. Missing data was inappropriately handled in two studies,[31,38] and was not reported in the remaining studies. As for model validation, every study, apart from one (that led to two articles),[40,41] used internal validation processes. Calibration was evaluated in two studies[33,36] using the McFadden R^2, isotonic regression and Brier score, and discrimination accuracy was assessed through AUC in five papers.[31,36,38-41] Reported AUCs ranged from 0.60 to 0.90. The model which showed apparent better discrimination ability was developed in the study by Xie and colleagues[36], using 37 clinical variables and a Light Gradient Boosting Machine (Light GBM). One study reported an accuracy of 94%[43], and one study[33] only reported an out-of-bag error rate, which varied from 31 to 63%. Kaskebar et al. and Husers et al. reported 95% CI, allowing comparisons between models.

Table 2. Characteristics of the included studies organized by model development stage, study design, setting and sample size: lower extremity amputation as outcome.

Reference	Study design, setting and follow-up	Study population and characteristics	Development/ validation	Variables assessed	Primary and secondary outcomes	Results	Comments
Stefanopoulos, 2022[31]	Retrospective cohort Multicenter National Inpatient Sample database from community hospitals in USA Inclusion period: 2008 to 2014 Follow-up period Mean LoS of 7.3 (±7.5) days on no LEA group and 12.9 (±10.5) days on LEA group	n = 326,853 people admitted with DFU Mean age: >62 years 36% male Mean diabetes duration: NR Exclusion criteria: NR	Development (train): CTREE with LASSO regression + random forest analysis 70% of sample randomly selected Internal validation (test): 30% of sample randomly selected + boosting (10 unique 5x boosted data sets)	39 variables tested, only 19 with OR provided: Gangrene, septic shock, peripheral vascular disease, weight loss, septicemia, systematic infection, anemia, age ≥ 40, bacteremia, elective procedure, osteomyelitis, leukocytosis, paralysis, renal failure, depression, male, congestive heart failure, hypertension, urinary tract infection, smoking Top 5 contributing variables (p<0.001): gangrene, osteomyelitis, systemic infection, peripheral vascular disease and weight loss	Major LEA	5.9% major LEA 5- and 10- variables model Unboosted and boosted Train Accuracy 78 % Sensitivity 76 % Specificity 79 % Test Accuracy 78 % Sensitivity 76 % Specificity 79 % AUC 0.84 Random forest Train and boosted Accuracy 78 % Sensitivity 76 % Specificity 80 % Boosted AUC 0.83	Retrospective Patient follow-up period NR Those with missing values or died during follow-up were excluded Only included risk factors with ICD-9 code entry 95% CI mostly not reported No external validation No calibration measures reported Final interactive model available on: https://grenut.shinyapps.io/amputation/
Austin, 2022[33]	Retrospective cohort Multicenter Fee-for-service Medicare database, USA Inclusion period: 2015 Follow-up period: 6 months	n = 88,898 people with DFU and PAD Mean age: 77 years 53 % female Mean diabetes duration: NR Inclusion criteria: newly diagnosed with concomitant PAD and diabetes; USA residents; age between 65 and 95; Medicare beneficiaries for 1 year following index date (first claim containing first diabetes-related ICD-9 or 10 code); ulcer diagnosis during first 6 months of the index year; outcome-free (alive, no reinterventions or amputations) for at least 6 months after index date	Development (train): 2/3 of sample Logistic regression; random forest, using a data set randomly divided into training (2/3) and testing (1/3) groups Calibration of logistic regression model: McFadden R^2 Internal validation (test) 1/3 of sample	9 variables tested Charlson comorbidity index, sex, race, age at diagnosis, Medicare-Medicaid dual-eligibility status, urban/rural indicator, HbA1c, foot exam, vascular imaging study	LEA (both minor and major)	LEA rate was 1.6% Logistic regression McFadden R^2: LEA: 0.071 Out-of-bag error rate: LEA: 63% Random forest Out-of-bag error rate: LEA: 31%	Retrospective Insurance based database No missing values reported Excluded lost to follow-up Information provided does not allow model application Not described how sample was split for train and test No external validation Calibration measures reported only for logistic regression model 95% CI not reported

Wang, 2022b[38]	Retrospective cohort Multicenter 2 tertiary hospitals in eastern China Inclusion period: January 2018 to December 2019 Follow-up period: NR	n = 362 people with DFU > 63 years 71% male >12 years duration Inclusion criteria: 18–89 years, University of Texas Grade 3 DFU Exclusion criteria: tumour-induced ulcers; previous major LEA; abandoned treatment; had incomplete information	Development (train): 70% of sample randomly selected 5 different models (logistic regression; random forest; decision tree; support vector machine; and XGBoost), using a data set randomly divided into training (70%) and testing (30%) groups; SMOTE method was used to solve the imbalance between non-LEA and minor LEA groups Internal validation (test): 30% of sample randomly selected; 10-fold cross-validation	21 variables assessed 9 variables selected by univariate analysis random blood glucose, years of diabetes, cardiovascular disease, peripheral arterial disease, smoking history, albumin, creatinine, CRP and DFU history	Minor LEA	20.7% minor LEA For minor LEA (after SMOTE): Decision tree Accuracy: 0.744 PPV: 0.828 Sensitivity: 0.616 F1 score: 0.707 AUC: 0.813 Random forest Accuracy: 0.797 PPV: 0.823 Sensitivity: 0.756 F1 score: 0.788 AUC: 0.857 Logistic regression Accuracy: 0.640 PPV: 0.640 Sensitivity: 0.640 F1 score: 0.640 AUC: 0.739 Support vector machine Accuracy: 0.663 PPV: 0.689 Sensitivity: 0.593 F1 score: 0.638 AUC: 0.767 XGBoost (model available at https://dfuprediction.azurewebsites.net/) Accuracy: 0.814 PPV: 0.846 Sensitivity: 0.767 F1 score: 0.805 AUC: 0.881	Retrospective Those lost to follow-up and with missing values were excluded 95% CI not reported No calibration measures reported No external validation

Xie, 2021[36]	Retrospective cohort Single-center University Hospital in China Inclusion period: 2009 to 2020 Follow-up period: NR	n = 618 people admitted with DFU Mean age: 66 years 62% male Mean diabetes duration: 10 years Exclusion criteria: <18 years; patients who died before hospital discharge; patients who underwent LEA before being referred to the hospital; patients who were referred to other medical institutions during the treatment period	Development (train): LightGBM 60% of sample randomly selected Calibration: Isotonic regression + Brier score Internal validation (test): 5-fold cross validation 20% of sample randomly selected SHAP algorithm for model explanation	37 variables included in the model: Age, sex, body mass index, diabetes duration, smoking history, pre-hospital delay, hypertension, coronary heart disease, heart failure, cerebral infarction, diabetic neuropathy, diabetic retinopathy, diabetic nephropathy, peripheral vascular disease, arterial occlusion, gangrene, prior DFU, prior amputation, HbA1C, random blood glucose, white blood cell, percentage of neutrophils, hemoglobin, serum potassium, serum sodium, serum creatinine, serum albumin, total cholesterol, triglyceride, LDL-C, HDL-C, antihyperglycemic drugs, insulin use, Wagner classification system, WIfI classification system (wound; ischemia; foot infection)	Minor LEA and major LEA	11.5% minor LEA, 7.6% major LEA Overall: AUC: 0.90 Sensitivity: 87.1% Specificity: 74.4% NPV: 79.7% PPV 86.3% Brier score: 0.086 Non-LEA: AUC: 0.90 Sensitivity: 95.0% Specificity: 69.6% NPV: 76.2% PPV 93.2% Minor LEA: AUC: 0.86 Sensitivity: 64.3% Specificity: 94.5% NPV: 95.4% PPV 60.0% Major LEA: AUC: 0.85 Sensitivity: 33.3% Specificity: 97.3% NPV: 94.9% PPV 50.0%	Retrospective Single-center study Patient follow-up period NR No missing values reported Information provided does not allow model application 95% CI not reported No external validation
Kasbekar, 2017[43]	Retrospective cohort Single-center Hospital in India Inclusion period: Jun 2011 to Jun 2013 Follow-up period: 1 year	n = 301 people admitted due to DFU Mean age: NR Male: NR Mean diabetes duration: NR Inclusion criteria: lower extremity ulcer, diabetes Exclusion criteria: those died, or lost to follow-up	Development (train): n=250 C5.0 (decision tree algorithm) Internal validation (test): n=51	21 variables: diabetes treatment, duration of leg symptoms, treatment for leg symptoms, history of regular smoking, history of regular alcohol intake, other comorbidities, age, sex, weight, Wagner classification, blood haemoglobin, serum creatinine, blood prothrombin time, serum albumin, serum bilirubin, random blood glucose level, HbA1C levels, x-ray of the affected foot, pus culture and sensitivity examination of the wound, flow through arterial Doppler, length of stay	LEA	27.6% LEA Single C5.0 Tree Decision tree If absent or monophasic peripheral flow → high risk of LEA If normal, triphasic, or biphasic grade DFU with Wagner Those with grade 4 and 5 → high risk of LEA Development Accuracy: 96% Validation Accuracy: 94% Boosted C5.0 Tree Ensemble (using 17 variables) Accuracy: 96% (95% CI: 0.865 – 0.995)	Retrospective Single-center study Baseline characteristics NR No missing values reported Those lost to follow-up were excluded No calibration measures reported No external validation

Du, 2021[39]	Retrospective cohort Single center University hospital in China Inclusion period: Pre-lockdown January to June 2019 and post-lockdown January to May 2020 Follow-up period: NR	n = 46 people admitted with DFU (23 + 23) Mean age: >66 years 76% male Mean diabetes duration: >11 years Inclusion criteria: DFU with IWGDF guidelines diagnostic criteria; WIfI grade 1-3, class 1-3 ischemia, and class 1-3 infection requiring emergency admission	Development (train) (n=10 without outcome and 3 with outcome randomly selected): 6 different models logistic regression support vector machine random forest GBDT artificial neural network XGBoost Internal validation (test) (n=8 without outcome and 2 with outcome randomly selected): 3-fold cross-validation	31 variables tested, being the variables with more weight Age, diabetes duration, diabetes treatment, nephropathy, education level, smoking pack-years, HDL, WBC, blood potassium, blood sodium, pre-hospital delay, WIfI wound, Ischemia	LEA	LEA not reported Logistic regression AUC: 0.76 Accuracy: 0.70 Sensitivity: 0.33 Specificity: 0.86 PPV: 0.50 NPV: 0.75 Support vector machine AUC: 0.60 Accuracy: 0.70 Sensitivity: 0.00 Specificity: 1.00 PPV: 0.00 NPV: 0.70 Random forest AUC: 0.67 Accuracy: 0.80 Sensitivity: 0.67 Specificity: 0.86 PPV: 0.67 NPV: 0.86 GBDT AUC: 0.67 Accuracy: 0.70 Sensitivity: 0.33 Specificity: 0.86 PPV: 0.50 NPV: 0.75 XGBoost AUC: 0.86 Accuracy: 0.80 Sensitivity: 0.67 Specificity: 0.86 PPV: 0.67 NPV: 0.86 Artificial neural network AUC: 0.71 Accuracy: 0.70 Sensitivity: 0.33 Specificity: 0.86 PPV: 0.50 NPV: 0.75	Retrospective Single-center study Patient follow-up period NR Small sample Models for LEA different between pre- and post-lockdown No missing values reported 95% CI not reported No calibration measures reported No external validation

Hüsers, 2020 and 2022[40,41]	Prospective cohort Single-center Wound care center in Germany Inclusion period: 1st June 2013 to 1st July 2019 Follow-up period: 6 months	n = 237 participants with DFU n=16 (6%) lost to follow up Mean age: 66 years 84% male Mean diabetes duration: NR Exclusion criteria: other wounds (e.g.: venous or arterial leg ulcers)	Development (train): Bayesian logistic regression model	7 variables included in the model Age, gender, PEDIS system classification composing variables (perfusion status, ulcer extent, ulcer depth, infection status and foot sensation)	Major LEA Any LEA	12.2% major LEA 31.6% LEA Major LEA: Cut off 0.28 (95% HDI 0.16 – 0.41) AUC: 0.80 (95% HDI 0.78 – 0.80) Sensitivity: 83% (95% HDI 71% – 93%) Specificity: 66% (95% HDI 53% – 77%) With no prior information: AUC: 0.765 (95% HDI 0.725–0.779) Any LEA, no prior information: AUC: 0.793 (95% HDI 0.778–0.801) Any and major LEA, prior information: AUC: 0.790 (95% HDI 0.774–0.802)	Single-center study No missing values reported Information provided does not allow model application Outcome defined using ICD-9 codes No calibration measures reported No validation

AUC: area under the curve; CI: confidence interval; DFU: diabetic foot ulcer; GBDT: gradient boosted decision trees; HbA1c: hemoglobin A1c; HDI: highest density intervals; HDL-C: high-density lipoprotein-cholesterol; ICD-9: International Classification of Diseases, 9th Revision; IWGDF: International Working Group on the Diabetic Foot; LASSO: least absolute shrinkage and selection operator; LDL-C: low-density lipoprotein-cholesterol; LEA: lower extremity amputation; LightGBM: light gradient-boosting machine; LoS: length of stay; NPV: negative predictive value; NR: not reported; OR: odds ratio; PAD: peripheral arterial disease; PEDIS: perfusion, extent, depth, infection, and sensation; PPV: positive predictive value; SHAP: shapley additive explanations; SMOTE: synthetic minority oversampling technique; USA: United States of America; WBC: white blood cell count; WIfI: wound, ischemia, and foot infection; XGBost: extreme gradient boosting.

Mortality

Mortality was defined as an outcome in two articles (15%)[33,39], and, in one of them[33], it was measured after six months (see Table 3). A total of 88,944 persons were included. The mortality rate varied from 4.5 to 17.4%.

The final models included a median of 10 variables per study, with Du et al.[39] only describing the most relevant variables for model construction. The only repeated variable was age (n=2, 100%) (see Supplementary Table 4).

Both studies used multiple ML techniques including logistic regression (LR) and RF in their analysis.

Overall, eight prediction models were created. Neither of the studies reported missing data. Both studies used processes of internal validation. Calibration was evaluated in one study using the McFadden R^2, and discrimination accuracy was assessed, by one paper[39], using the AUC, with reported values ranging from 0.56 to 0.94. The other paper[33] only reported out-of-bag error rate varying from 30 to 68%.

Reintervention

The necessity of reintervention was used as an outcome in one study,[33] and was measured after six months (see Table 4). A total of 88,898 subjects were included in this study, presenting a reintervention rate of 3.6%.

Two prediction models were created using nine variables, using LR and RF. Missing data was not reported, and internal validation was used. Calibration was assessed using the McFadden R^2. Out-of-bag error rate varied from 36 (RF) to 60% (LR).

Table 3. Characteristics of the included studies organized by model development stage, study design, setting and sample size: mortality as outcome.

Reference	Study design, setting and follow-up	Study population and characteristics	Development/ validation	Variables assessed	Primary and secondary outcomes	Results	Comments
Austin, 2022[33]	Retrospective cohort Multicenter Fee-for-service Medicare database, USA Inclusion period: 2015 Follow-up period: 6 months	n = 88,898 people with DFU and PAD Mean age: 77 years 53 % female Mean diabetes duration: NR Inclusion criteria: newly diagnosed with concomitant PAD and diabetes; USA residents; age between 65 and 95; Medicare beneficiaries for 1 year following index date (first claim containing first diabetes-related ICD-9 or 10 code); ulcer diagnosis during first 6 months of the index year; outcome-free (alive, no reinterventions or amputations) for at least 6 months after index date	Development (train): 2/3 of sample Logistic regression; random forest, using a data set randomly divided into training (2/3) and testing (1/3) groups Calibration of logistic regression model: McFadden R^2 Internal validation (test) 1/3 of sample	9 variables tested Charlson comorbidity index, sex, race, age at diagnosis, Medicare-Medicaid dual-eligibility status, urban/rural indicator, HbA1c, foot exam, vascular imaging study	Death	Mortality rate was 4.5% Logistic regression McFadden R^2:0.046 Out-of-bag error rate:68% Random forest Out-of-bag error rate: 30%	Retrospective Insurance based database No missing values reported Excluded lost to follow-up Information provided does not allow model application Not described how sample was split for train and test No external validation 95% CI not reported
Du, 2021[39]	Retrospective cohort Single center University hospital in China Inclusion period: Pre-lockdown January to June 2019 and post-lockdown January to May 2020 Follow-up period: NR	n = 46 people admitted with DFU (23 + 23) Mean age: >66 years 76% male Mean diabetes duration: >11 years Inclusion criteria: DFU with IWGDF guidelines diagnostic criteria; WIfI grade 1-3, class 1-3 ischemia, and class 1-3 infection requiring emergency admission	Development (train) (n=10 without outcome and 3 with outcome randomly selected): 6 different models logistic regression support vector machine random forest GBDT artificial neural network XGBoost Internal validation (test) (n=8 without outcome and 2 with outcome randomly selected): 3-fold cross-validation	31 variables tested, being the variables with more weight For mortality: Infection (not foot related), age, foot infection, PAD, WBC, education level, diabetes treatment, LDL, cerebral infarction, HDL	Mortality	Mortality 0% pre- vs. 17.4% post-lockdown Logistic regression AUC: 0.81 Accuracy: 0.70 Sensitivity: 0.50 Specificity: 0.75 PPV: 0.33 NPV: 0.86 Support vector machine AUC: 0.56 Accuracy: 0.80 Sensitivity: 0.00 Specificity: 1.00 PPV: 0.00 NPV: 0.80 Random forest AUC: 0.88 Accuracy: 0.80 Sensitivity: 0.50 Specificity: 0.88 PPV: 0.50 NPV: 0.88	Retrospective Single-center study Patient follow-up period NR Small sample No missing values reported 95% CI not reported No calibration measures reported No external validation

			GBDT AUC: 0.88 Accuracy: 0.90 Sensitivity: 1.00 Specificity: 0.88 PPV: 0.67 NPV: 1.00 XGBoost AUC: 0.94 Accuracy: 0.90 Sensitivity: 1.00 Specificity: 0.88 PPV: 0.67 NPV: 1.00 Artificial neural network AUC: 0.69 Accuracy: 0.70 Sensitivity: 0.50 Specificity: 0.75 PPV: 0.33 NPV: 0.86

AUC: area under the curve; CI: confidence interval; DFU: diabetic foot ulcer; GBDT: gradient boosted decision trees; HbA1c: hemoglobin A1c; HDL-C: high-density lipoprotein-cholesterol; ICD-9: International Classification of Diseases, 9th Revision; IWGDF: International Working Group on the Diabetic Foot; LDL-C: low-density lipoprotein-cholesterol; NPV: negative predictive value; NR: not reported; PAD: peripheral arterial disease; PPV: positive predictive value; USA: United States of America; WBC: white blood cells count; WIfI: wound, ischemia, and foot infection; XGBost: extreme gradient boosting.

Table 4. Characteristics of the included studies organized by model development stage, study design, setting and sample size: reintervention as outcome.

Reference	Study design, setting and follow-up	Study population and characteristics	Development/ validation	Variables assessed	Primary and secondary outcomes	Results	Comments
Austin, 2022[28]	Retrospective cohort Multicenter Fee-for-service Medicare database, USA Inclusion period: 2015 Follow-up period: 6 months	n = 88,898 people with DFU and PAD Mean age: 77 years 53 % female Mean diabetes duration: NR Inclusion criteria: newly diagnosed with concomitant PAD and diabetes; USA residents; age between 65 and 95; Medicare beneficiaries for 1 year following index date (first claim containing first diabetes-related ICD-9 or 10 code); ulcer diagnosis during first 6 months of the index year; outcome-free (alive, no reinterventions or amputations) for at least 6 months after index date	Development (train): 2/3 of sample Logistic regression; random forest, using a data set randomly divided into training (2/3) and testing (1/3) groups Calibration of logistic regression model: McFadden R^2 Internal validation (test) 1/3 of sample	9 variables tested Charlson comorbidity index, sex, race, age at diagnosis, Medicare-Medicaid dual-eligibility status, urban/rural indicator, HbA1c, foot exam, vascular imaging study	Reintervention	Reintervention rate was 3.6% Logistic regression McFadden R^2:0.124 Out-of-bag error rate:60% Random forest Out-of-bag error rate: 36%	Retrospective Insurance based database No missing values reported Excluded lost to follow-up Information provided does not allow model application Not described how sample was split for train and test No external validation 95% CI not reported

CI: confidence interval; DFU: diabetic foot ulcer; HbA1c: hemoglobin A1c; ICD-9: International Classification of Diseases, 9th Revision; NR: not reported; PAD: peripheral arterial disease; USA: United States of America.

Risk of bias

We evaluated the risk of bias according to two tools: QUIPS (see Table 5), with a proportion of agreement of 71%, and PROBAST (see Table 6), with a proportion of agreement of 88% for risk of bias and 61% for applicability.

Risk of bias according to QUIPS

All studies presented a high risk of bias and had two or three (out of the six domains) classified as being at low risk (see Table 5).

In the study participation domain, all 11 studies were classified as having a moderate or high risk of bias, mainly because most studies did not clearly explain how participants' sampling was conducted and failed to describe fully eligibility criteria.

In the study attrition domain, nine studies (82%) were classified as having a high risk of bias. Of these, seven did not mention the proportion of patients that concluded the study[31,33,36-41,43], and two[34,35], although reported the response rate, did not characterise the excluded patients or the reasons for their exclusion.

As for the prognostic factor measurement domain, all studies were classified as having moderate or high risk of bias. Nine studies (82%)[31,32,34-39,42,43] failed to clearly define or explain how to collect all the analysed variables. Several variables, such as wound area, wound depth, erythema and adequate arterial flow, are subjective and can lead to different results if measured differently. Two studies[32,35] may have introduced bias due to using several patient centres with no standardised protocols provided. Apart from one study[34], in which imputation with a k-NN algorithm was used, missing data was not reported, not addressed or incorrectly handled.

Concerning the outcome measurement domain, four studies (36%) were classified as having a moderate or high risk of bias. Three of these studies[32,34,35] failed to define the outcome clearly, and none specified a protocol for outcome assessment.

Statistical analysis and reporting were considered in all studies to be at low risk of bias.

Risk of bias according to PROBAST

All studies presented a high risk of bias, and 10 presented high or unclear concerns for applicability (see Table 6).

Five studies were classified as having high or unclear risk of bias in the participants' domain due to inadequate or absent eligibility criteria. In the evaluation of applicability, seven studies had a high or unclear concern for applicability. Similarly, this is mainly due to the inclusion of participants that may differ from the studies' target population.

In the predictors domain, nine studies were classified as having unclear risk of bias and unclear risk of applicability. Eight studies[31,33-39,43] had a retrospective design and failed to mention if the predictors'

assessment was made without the knowledge of outcome data. Two studies[32,35] included multiple patient centres, and a homogenous assessment of predictors was not clearly guaranteed.

As for the outcomes, nine papers were classified as having a high or unclear risk of bias. In four of these[32,34,35,42], authors failed to report a clear definition for the outcome, compromising its measurement. Another factor that potentially introduced bias was the lack of follow-up reporting (n=5)[31,36,38,39,43]. In the evaluation of applicability, four studies had a high concern for applicability due to the lack of clear definitions for the outcomes.

In the analysis domain, all studies were classified as having a high risk of bias. Nine studies had insufficient sample sizes, with only two studies[31,33] reaching over 200 events per variable (EPV) tested – the cut-off considered reasonable to minimise overfitting when using ML techniques. Three studies[31,35,37,38] converted continuous variables into categorical ones, using arbitrary rules for categorisation. Regarding missing data, only one study[34] addressed missing data correctly, using imputation with a k-NN algorithm. Two studies[37,38,42] failed to avoid selecting variables solely based on univariate analysis, whereas two[35,43] did not mention how variables were selected. As for model evaluation, seven studies[31,34,37-43] did not report any calibration. The remaining used different methods, including Hosmer-Lemeshow statistics, the Brier score, the McFadden R^2 and the isotonic regression. Discrimination accuracy measures were employed by all studies except one[33], with AUC being the most often used. However, uncertainty measures were seldom provided. In terms of validation, apart from one (that only derived models)[40,41], every study relied solely on internal validation methods.

Table 5. QUIPS tool results of included studies.

Author, year	Study participation	Study attrition	Prognostic factor measurement	Outcome measurement	Study confounding	Statistical analysis and reporting	Overall RoB
Austin, 2022[28]	Moderate	High	Moderate	Low	Low	Low	High
Du, 2021[34]	Moderate	High	High	Low	Low	Low	High
Hüsers, 2020 and 2022[35,36]	Moderate	High	Moderate	Low	Low	Low	High
Jung, 2016[30]	Moderate	High	High	High	Low	Low	High
Kasbekar, 2017[38]	High	High	High	Low	Low	Low	High
Kim, 2020[29]	Moderate	High	Moderate	Moderate	Low	Low	High
Margolis, 2022[27]	Moderate	Low	High	Moderate	Low	Low	High
Poradzka, 2023[37]	Moderate	Low	High	Moderate	Low	Low	High
Stefanopoulos, 2022[26]	Moderate	High	High	Low	Low	Low	High
Wang, 2022a and b[32,33]	Moderate	High	Moderate	Low	Low	Low	High
Xie, 2021[31]	High	High	High	Low	Low	Low	High

RoB: risk of bias. A darker tone of the cell represents a higher risk of bias.

Table 6. PROBAST tool results of included studies.

Author, year	RoB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	RoB	Applicability
Austin, 2022[28]	+	?	+	-	-	?	+	-	-
Du, 2021[34]	+	?	?	-	-	?	+	-	-
Hüsers, 2020 and 2022[35,36]	+	+	+	-	+	+	+	-	+
Jung, 2016[30]	?	?	?	-	?	?	-	-	-
Kasbekar, 2017[38]	+	?	?	-	+	?	+	-	?
Kim, 2020[29]	-	?	-	-	-	?	-	-	-
Margolis, 2022[27]	?	?	-	-	-	?	-	-	-
Poradzka, 2023[37]	+	+	-	-	+	+	-	-	-
Stefanopoulos, 2022[26]	?	?	?	-	?	?	+	-	?
Wang, 2022a[32]	-	?	+	-	-	?	+	-	-
Wang, 2022b[33]			?						
Xie, 2021[31]	+	?	?	-	+	?	+	-	?

RoB: risk of bias; +: low RoB or low concern regarding applicability; -: high RoB or high concern regarding applicability; ?: unclear RoB or unclear concern regarding applicability. A darker tone of the cell represents a higher risk of bias.

Discussion

DFUs have become one of the most important causes of mortality and morbidity in people with diabetes. Guideline-directed foot exams and treatments and aggressive medical management of diabetes and cardiovascular disease are paramount in improving the prognosis of people with a DFU. However, effective interventions do not work in the same way in people, and the same management cannot be provided to all people with diabetes. For that reason, predictive models have been used to stratify people with DFU by their probability of healing, requiring an LEA and dying, so that interventions and resources can be appropriately allocated.

Our systematic review aimed to ascertain if models using an ML approach and clinical data that are easily accessible in clinical practice could predict clinical outcomes. We have included 11 studies corresponding to 13 articles, mainly from the USA and China (73%), retrospective (73%), single centre (55%), and LEA as the outcome (54%).

Other reviews have previously investigated how ML could improve DFU care. For example, a systematic review by Tulloch and colleagues[23] focused on identifying and classifying DFU at a specific moment. In this review, the predictive abilities of ML algorithms were not considered, unlike ours. More recently, a systematic review by Huang and colleagues[44] searched the literature for models that predicted amputation. This study focused on prognosis but considered models that used several methodologies (ML and not ML) and that predicted only one outcome (LEA). This review included people with "diabetic foot", a broader concept than DFU, characterised by ulcers or destruction of the tissues due to infection and/or peripheral artery disease.

Although we have included 11 studies (published as 13 articles), we found a total of 57 prediction models focusing on four outcomes: healing (5 articles presented 20 models), LEA (7 articles presented 27 models), mortality (2 articles presented eight models) and reintervention (1 study presented two models). Our search did not retrieve some of the outcomes defined in our protocol, such as hospitalisation, length of stay, health-related quality of life, ulcer-free survival period/time and LEA-free period, nor reliability studies, external validation studies, nor studies assessing the impact of developing and implementing a predictive model in clinical practice.

For model construction, studies used clinical variables distributed into four categories: demographic characteristics, medical history, laboratory data and foot-related characteristics. The number of variables in the final models varied from four to 865.

When it comes to wound healing, variables from the foot-related characteristics group were the most frequently selected, with wound area being used in models from almost all studies. In fact, in previous studies, ischemic ulcers, more extensive and deeper ulcers, plantar ulcers, and ulcers with infection have been associated with poor healing.[45,46] Besides clinical variables, Kim et al.[34] also included image-based characteristics from photographs through subjective and deep learning analysis. The subjective

observation allowed models to be adequately trained with good performance, turning the utilisation of smartphone and tablet photographs for prognosis assessment in clinical practice into a possibility.

Regarding LEA, the most frequently selected variables were age, gender, diabetes duration, smoking history, haemoglobin A1c, creatinine, albumin and random blood glucose. In the case of gender, males have been reported to have higher amputation rates, likely due to behavioural differences.[47,48] As for smoking, studies have suggested its association with LEA by increasing the risk of atherosclerosis and, consequently, of PAD.[47,48] Decreased albumin levels have also been connected to higher LEA rates.[48] As for the remaining variables, a recent systematic review with meta-analysis found no correlation between these and the risk of LEA[48], questioning the methods chosen by these papers to select predictors for model construction. Lastly, in the case of mortality, age was the only variable repeated in the two selected studies. In the case of reintervention, only one study was selected by our systematic review. Despite these being two important outcomes, not many studies have sought to create models that could predict mortality or reintervention.

The most used ML method was RF. It was first described by Breiman et al.[49], and is a supervised method that uses "parallel ensembling", fitting several decision tree classifiers in parallel, where each tree is trained on a random subset of the training data with replacement (bagging). Majority voting or averages are used to obtain the final result. This method is suitable for both classification and regression problems and has a reduced risk of overfitting, when compared to decision trees. However, it is a time-consuming process, that requires more resources, and with a more complex interpretation.[21]

The most reported discrimination accuracy measure was AUC. It refers to the area under the receiver operating characteristic (ROC) curve and is an effective way of summarising the overall diagnostic accuracy of a model, taking values from zero to one. The ROC curve is depicted by using each possible value of a continuous variable as a point with a certain sensitivity and (1-) specificity in discriminating those with and without the clinical condition of interest.[50]

There is some variation in the qualitative descriptors of model performance for AUC thresholds. According to Mandrekar et al.[50], a AUC value of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding. However, the AUC is a combined measure of the overall sensitivity and specificity of a model. This implies that two models can have identical AUC values, and one perform better in higher sensitivities and the other perform better in higher specificities.[51] Therefore, interpretation of AUC values must be done cautiously.

Overall, most models were able to achieve good discrimination ability, with 51% of the reported AUC values being equal or superior to 0.8. However, all studies were considered to have a high risk of bias, according to QUIPS and PROBAST. Firstly, most studies failed to clearly describe inclusion and/or exclusion criteria, raising doubt about the potential applicability of final models. Patients lost to follow-up, when reported, were also excluded. Secondly, most studies did not clearly define the variables considered or describe the methodology adopted to measure them. Some variables, such as those in the foot-related characteristics group, were subjective, and different assessors may measure them in

various ways. As a result, without clear definitions and established protocols for variables' measurement, predictive models' validity and application to clinical practice may be compromised. This lack of standardisation also applies to outcomes, where clear definitions are essential to guarantee consistency. In the case of the studies considering wound healing as an outcome, different variations were used (delayed wound healing, hard-to-heal wound, and wound healing failure). However, most did not provide a definition or a methodology for outcome measurement. Also, follow-up periods varied widely among studies, which makes comparing predictive models even more challenging. These variations and lack of standardisation were expected; thus, no meta-analysis was considered from inception. Thirdly, there were some problems in the analysis domain. Studies had insufficient sample sizes, with only two reaching over 200 EPV. 'EPV' refers to the number of events (i.e., number of patients in which the outcome of interest has occurred) relative to the number of regression coefficients used (i.e., number of variables considered for model development).[30] In ML-based models, higher EPV (often more than 200) are needed to minimize overfitting.[52] Missing data was not reported, not addressed or incorrectly handled, except in one study that used imputation methods.

Although reported models showed promise in predicting clinical outcomes, most are not available for immediate applicability to clinical practice. Only three studies presented online interactive models: two[31,38] for LEA prediction and one[37] for hard-to-heal DFU prediction. In addition, several studies developed and internally validated several models using different methods in the same sample without providing all required accuracy measures (focusing on AUC) and seldomly reporting 95% confidence intervals – making it impossible to compare and select specific models to be used in clinical practice.

Our systematic review presents some limitations. Firstly, we conducted our search in two phases, involving different reviewers in the selection process, which may have introduced some variability. However, the agreement proportion was high and comparable when comparing both phases. Secondly, a systematic review is always as good as the included studies. Due to the high risk of bias and the high heterogeneity of included studies (mainly due to a lack of uniform variable definitions, outcome definitions and follow-up periods), a meta-analysis was not adequate to perform. However, we would like to highlight that this reality is not as different from the results achieved in our systematic review of "classic" classification systems. Lastly, the complexity of the ML methodology prevented further explanation of the studied models, which may lead to some distrust from healthcare providers when considering the application of such models in clinical practice.

For future research, studies with larger samples with a low risk of bias, preferably in an external validation setting, are needed. When developing DFU prediction models, we recommend that variables be selected according to the available evidence provided by previous studies. Additionally, other outcomes, such as quality of life, which are clinically relevant, should be further utilised.

Conclusion

This systematic review found several ML-based models that could predict clinical outcomes with good discrimination ability in people with DFU, showing promising results. However, studies presented a high risk of bias with several applicability issues that compromise the ready applicability of such models in clinical practice. Further studies with stricter methodology are needed so that patients who have diabetes and a foot ulcer can benefit from the recent advancements of artificial intelligence applied to healthcare.

Supplementary material

Supplementary Table 1. Search queries on Pubmed, Web of Science, SCOPUS and IEEE Xplore.

Pubmed	("deep learning" [MeSH Terms] OR "deep learning" [All Fields] OR "artificial intelligence" [MeSH Terms] OR "artificial intelligence" [All Fields] OR "deep learning" [MeSH Terms] OR "deep learning" [All Fields] OR "random forest" [All Fields] OR "clinical decision-making"[MeSH Terms] OR "clinical decision-making"[All Fields] OR "medical decision-making"[All Fields] OR "decision support systems, clinical"[MeSH Terms] OR "decision support systems clinical"[All Fields] OR "clinical decision support systems"[All Fields] OR "decision support techniques"[MeSH Terms] OR "decision support techniques"[All Fields] OR "machine learning"[MeSH Terms] OR "machine learning"[All Fields] OR "data mining"[MeSH Terms] OR "data mining"[All Fields] OR "neural networks, computer"[MeSH Terms] OR "neural networks computer"[All Fields] OR "computer neural network*"[All Fields] OR "neural network*"[All Fields] OR ("phenotype"[MeSH Terms] OR "phenotype"[All Fields] OR "phenotypes"[All Fields] OR "phenotyped"[All Fields] OR "phenotypic"[All Fields] OR "phenotypical"[All Fields] OR "phenotypically"[All Fields] OR "phenotyping"[All Fields] OR "phenotypings"[All Fields]) OR "convolutional"[All Fields] OR automat* [Title/Abstract]) AND ("Diabetic Foot"[MeSH Terms] OR ("diabete"[All Fields] OR "diabetes mellitus"[MeSH Terms] OR ("diabetes"[All Fields] AND "mellitus"[All Fields]) OR "diabetes mellitus"[All Fields] OR "diabetes"[All Fields] OR "diabetic"[All Fields] OR "diabetics"[All Fields] OR "diabets"[All Fields]) AND ("ulcer"[MeSH Terms] OR "ulcer"[All Fields] OR "ulcerate"[All Fields] OR "ulcerated"[All Fields] OR "ulcerates"[All Fields] OR "ulcerating"[All Fields] OR "ulceration"[All Fields] OR "ulcerations"[All Fields] OR "ulcerative"[All Fields] OR "ulcers"[All Fields] OR "ulcerous"[All Fields]) OR ("amputate"[All Fields] OR "amputated"[All Fields] OR "amputating"[All Fields] OR "amputation"[MeSH Terms] OR "amputation"[All Fields] OR "amputations"[All Fields] OR "amputed"[All Fields]))
Web of Science	(TI= "deep learning" OR AB="deep learning" OR TI="artificial intelligence" OR AB="artificial intelligence" OR TI= "random forest" OR AB="random forest" OR TI="decision making" OR AB="decision making" OR TI="decision support" OR AB="decision support" OR TI="machine learning" OR AB="machine learning" OR TI="data mining" OR AB="data mining" OR TI="neural network" OR AB="neural network" OR TI="phenotype" OR AB="phenotype" OR TI="convolutional" OR AB="convolutional") AND ((TI="foot" OR AB="foot") AND (TI="ulcer" OR AB="ulcer"))
SCOPUS	(TITLE-ABS-KEY (deep AND learning) OR TITLE-ABS-KEY (artificial AND intelligence) OR TITLE-ABS-KEY (random AND forest) OR TITLE-ABS-KEY (decision AND making) OR TITLE-ABS-KEY (decision AND support) OR TITLE-ABS-KEY (machine AND learning) OR TITLE-ABS-KEY (data AND mining) OR TITLE-ABS-KEY (neural AND network) OR TITLE-ABS-KEY (phenotype) OR TITLE-ABS-KEY (convolutional)) AND ((TITLE-ABS-KEY (foot) AND TITLE-ABS-KEY (ulcer))
IEEE Xplore	("All Metadata":diabetic foot ulcer) AND ("All Metadata":classification) ("All Metadata":wound) AND ("All Metadata":chronic) AND ("All Metadata":classification) ("All Metadata":diabetic foot) AND ("All Metadata":healing) ("All Metadata":diabetic foot) AND ("All Metadata":convolutional) ("All Metadata":diabetic foot) AND ("All Metadata":amputation)

Supplementary Table 2. Distribution of clinical variables included in the final models by categories having healing as outcome.

Variable categories		Variables	Studies (References)
Demographic characteristics		Gender	3 (34, 35, 37)
		Age	2 (34,35)
		Center code	1 (35)
		Ethnicity/race	1 (34)
		Primary insurance	1 (35)
Medical history	Comorbidities	Diabetes-related retinopathy	2 (34,37)
		Charlson comorbidity index	1 (34)
		CKD stage	1 (34)
		History of amputation	1 (42)
		Palliative care	1 (35)
		Peripheral arterial disease	1 (37)
		Smoking history	1 (37)
	Drugs	Oral antihyperglycemic drugs/insulin/canagliflozin	2 (34,37)
		Allopurinol	1 (34)
		Antihypertensives	1 (34)
		Aspirin	1 (34)
		Heparin/warfarin/Xa inhibitors	1 (34)
		Immunosuppressants/oral steroids	1 (34)
		NSAIDs	1 (34)
	Others	Body mass index	2 (32,34)
		Body surface area	1 (34)
Laboratory data		CRP	3 (34,37,42)
		Albumin	2 (34,37)
		ESR	2 (34,42)
		Creatinine	1 (37)
		HbA1c	1 (34)
		Lymphocyte count	1 (34)
		Pre-albumin	1 (34)
		Random blood glucose	1 (37)
Foot related characteristics		Wound area	4 (32,35,37,42)
		Probe-to-bone test	2 (34,42)
		Temperature	2 (35,42)
		Wound depth	2 (34,35)
		Wound duration	2 (32,42)
		Wound length	2 (34,35)
		Wound width	2 (34,35)
		ABI	1 (42)
		Ankle systolic pressure measured/ toe systolic pressure measured	1 (34)
		Atrophie blanche	1 (35)
		Brawny induration	1 (35)
		Callus	1 (35)
		Clustered wound	1 (35)
		Complex care	1 (35)
		Crepitus	1 (35)
		Cyanosis	1 (35)
		Delayed recurrence	1 (35)
		Dorsalis pedis pulse measured/ posterior tibial pulse measured	1 (34)
		Duration of care	1 (35)
		Ecchymosis	1 (35)
		Edema	1 (35)
		Epithelization	1 (35)
		Erythema	1 (35)
		Excoriation	1 (35)
		Exposed bone	1 (35)
		Exposed joint	1 (35)
		Exposed muscle	1 (35)
		Exposed tendon	1 (35)
		Exudate	1 (35)

	Fluctuance	1 (35)
	Foot	1 (34)
	Friable	1 (35)
	Granulation amount	1 (35)
	Granulation quality	1 (35)
	Hemosiderosis	1 (35)
	Infection	1 (34)
	Margin	1 (35)
	MRI performed	1 (34)
	Number of wounds	1 (35)
	Offload	1 (34)
	Pallor	1 (35)
	Pending amputation on presentation	1 (35)
	Presence of blood flow in the Doppler probe	1 (42)
	Previous wound count	1 (35)
	Rash	1 (35)
	Result of accident	1 (35)
	Rubor	1 (35)
	Slough	1 (35)
	Tc99 bone scan performed	1 (34)
	Tenderness on palpation	1 (35)
	Thickness	1 (35)
	Total contact cast use	1 (34)
	Transcutaneous oxygen pressure	1 (34)
	Tunneling	1 (35)
	Undermining	1 (35)
	UTSA stage/ UTSA grade (ischemia)/ wound grade	1 (34)
	Wound location	1 (35)
	Wound recurrence	1 (35)
	Wound type	1 (35)
	X-ray performed	1 (34)

ABI: ankle-brachial index; CKD: chronic kidney disease; CRP: C-reactive protein; ESV: erythrocyte sedimentation rate; HbA1c: hemoglobin A1c; MRI: magnetic resonance imaging; NSAIDs: non-steroidal anti-inflammatory drugs; Tc99: technetium-99 bone scan; UTSA: University of Texas Santo Antonio staging.

Supplementary Table 3. Distribution of clinical variables included in the final models by categories having lower extremity amputation as outcome.

Variable categories		Variables	Studies (References)
Demographic characteristics		Age	5 (33,36,39,40,41,43)
		Gender	3 (33,36,40,41)
		Education level	1 (39)
		Insurance status	1 (33)
		Race	1 (33)
		Urban/rural	1 (33)
Medical history	Comorbidities	Diabetes duration	3 (36,38,39)
		Smoking history	3 (36,38,39)
		DFU history	2 (36,38)
		Diabetes-related nephropathy	2 (36,39)
		Peripheral vascular disease	2 (31,36)
		Arterial occlusion	1 (36)
		Cardiovascular disease	1 (38)
		Cerebral infarction	1 (36)
		Charlson comorbidity index	1 (33)
		Coronary heart disease	1 (36)
		Diabetic neuropathy	1 (36)
		Diabetic retinopathy	1 (36)
		Heart failure	1 (36)
		History of diabetes	1 (43)
		Hypertension	1 (36)
		Other comorbidities	1 (43)
		Peripheral arterial disease	1 (38)
		Prior amputation	1 (36)
		Systemic infection	1 (31)
	Drugs	Antihyperglycemic drugs/insulin	2 (36,39)
	Others	Pre-hospital delay	2 (36,39)
		Weight/weight loss	2 (31,43)
		Body mass index	1 (36)
		Days admitted	1 (43)
		Duration of symptoms	1 (43)
Laboratory data		Albumin	3 (36,38,43)
		Creatinine	3 (36,38,43)
		HbA1c	3 (33,36,43)
		Random blood glucose	3 (36,38,43)
		HDL	2 (36,39)
		Hemoglobin	2 (36,43)
		Potassium	2 (36,39)
		Sodium	2 (36,39)
		WBC/percentage of neutrophils	2 (36,38)
		Bilirubin	1 (43)
		CRP	1 (38)
		Culture report growth	1 (43)
		LDL	1 (36)
		Prothrombin time	1 (43)
		Total cholesterol	1 (36)
		Triglyceride	1 (36)
Foot related characteristics		Gangrene	2 (31,36)
		Vascular imaging study/arterial doppler flow	2 (33,43)
		WIfI classification system	2 (36,38)
		Foot exam	1 (33)
		Foot X-ray	1 (43)
		Ischemia	1 (39)
		Osteomyelitis	1 (31)
		PEDIS classification system	1 (40,41)
		Ulcer grade	1 (43)
		Wagner classification system	1 (36)

CRP: C-reactive protein; DFU: diabetic foot ulcer; HbA1c: hemoglobin A1c; HDL: high-density lipoprotein; LDL: low-density lipoproteins; PEDIS: perfusion status, ulcer extent, ulcer depth, infection status and foot sensation; WBC: white blood cells count; WiFI: wound, ischemia, and foot infection.

Supplementary Table 4. Distribution of clinical variables included in the final models by categories having mortality as outcome.

Variable categories		Variables	Studies (References)
Demographic characteristics		Age	2 (33,39)
		Education level	1 (39)
		Gender	1 (33)
		Insurance status	1 (33)
		Race	1 (33)
		Urban/rural	1 (33)
Medical history	Comorbidities	Cerebral infarction	1 (39)
		Charlson comorbidity index	1 (33)
		Infection (not foot related)	1 (39)
		Peripheral arterial disease	1 (39)
	Drugs	Diabetes treatment	1 (39)
Laboratory data		HbA1c	1 (33)
		HDL	1 (39)
		LDL	1 (39)
		WBC	1 (39)
Foot related characteristics		Foot exam	1 (33)
		Foot infection	1 (39)
		Vascular imaging study	1 (33)

HbA1c: hemoglobin A1c; HDL: high-density lipoprotein; LDL: low-density lipoprotein; WBC: white blood cells count.

References

1. Sociedade Portuguesa de Diabetologia. Diabetes: Factos e Números – O Ano de 2019, 2020 e 2021 – Relatório Anual do Observatório Nacional da Diabetes 03/2023. 2023.
2. Direção-Geral da Saúde. Programa Nacional para a Diabetes: Desafios e Estratégias 2023. 2023.
3. Reardon R, Simring D, Kim B, Mortensen J, Williams D, Leslie A. The diabetic foot ulcer. Australian Journal for General Practitioners 2020;49:250-255. (https://www1.racgp.org.au/ajgp/2020/may/diabetic-foot-ulcer).
4. Musuuza J, Sutherland BL, Kurter S, Balasubramanian P, Bartels CM, Brennan MB. A systematic review of multidisciplinary teams to reduce major amputations for patients with diabetic foot ulcers. Journal of Vascular Surgery 2020;71(4):1433-1446.e3. DOI: 10.1016/j.jvs.2019.08.244.
5. Monteiro-Soares M, Hamilton EJ, Russell DA, et al. Guidelines on the classification of foot ulcers in people with diabetes (IWGDF 2023 update). Diabetes Metab Res Rev 2023:e3648. DOI: 10.1002/dmrr.3648.
6. Young RA, Martin CM, Sturmberg JP, et al. What Complexity Science Predicts About the Potential of Artificial Intelligence/Machine Learning to Improve Primary Care. The Journal of the American Board of Family Medicine 2024;37(2):332. DOI: 10.3122/jabfm.2023.230219R1.
7. International Diabetes Federation. IDF Diabetes Atlas. 2021.
8. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: a review of current evidence. Diabetologia 2019;62(1):3-16. (In eng). DOI: 10.1007/s00125-018-4711-2.
9. van Netten JJ, Bus SA, Apelqvist J, et al. Definitions and criteria for diabetic foot disease. Diabetes Metab Res Rev 2020;36 Suppl 1:e3268. DOI: 10.1002/dmrr.3268.
10. Armstrong DG, Boulton AJM, Bus SA. Diabetic Foot Ulcers and Their Recurrence. N Engl J Med 2017;376(24):2367-2375. DOI: 10.1056/NEJMra1615439.
11. Hoffstad O, Mitra N, Walsh J, Margolis DJ. Diabetes, lower-extremity amputation, and death. Diabetes Care 2015;38(10):1852-7. DOI: 10.2337/dc15-0536.
12. Meloni M, Izzo V, Giurato L, Lazaro-Martinez JL, Uccioli L. Prevalence, Clinical Aspects and Outcomes in a Large Cohort of Persons with Diabetic Foot Disease: Comparison between Neuropathic and Ischemic Ulcers. J Clin Med 2020;9(6). DOI: 10.3390/jcm9061780.
13. 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):209-221. DOI: 10.2337/dci22-0043.
14. Khunkaew S, Fernandez R, Sim J. Health-related quality of life among adults living with diabetic foot ulcers: a meta-analysis. Qual Life Res 2019;28(6):1413-1427. (In eng). DOI: 10.1007/s11136-018-2082-2.
15. Monteiro-Soares M, Hamilton EJ, Russell DA, et al. Classification of foot ulcers in people with diabetes: A systematic review. Diabetes Metab Res Rev 2024;40(3):e3645. DOI: 10.1002/dmrr.3645.
16. Won SH, Chung CY, Park MS, et al. Risk factors associated with amputation-free survival in patient with diabetic foot ulcers. Yonsei Med J 2014;55(5):1373-8. DOI: 10.3349/ymj.2014.55.5.1373.
17. Lu Q, Wang J, Wei X, Wang G, Xu Y. Risk Factors for Major Amputation in Diabetic Foot Ulcer Patients. Diabetes Metab Syndr Obes 2021;14:2019-2027. DOI: 10.2147/DMSO.S307815.
18. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart

Failure Outcomes. JAMA Network Open 2020;3(1):e1918962-e1918962. DOI: 10.1001/jamanetworkopen.2019.18962.

19. Premsagar P, Aldous C, Esterhuizen TM, Gomes BJ, Gaskell JW, Tabb DL. Comparing conventional statistical models and machine learning in a small cohort of South African cardiac patients. Informatics in Medicine Unlocked 2022;34:101103. DOI: https://doi.org/10.1016/j.imu.2022.101103.
20. Mohammed M, Khan M, Bashier E. Machine Learning: Algorithms and Applications2016.
21. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2021;2(3):160. DOI: 10.1007/s42979-021-00592-x.
22. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017;15:104-116. DOI: 10.1016/j.csbj.2016.12.005.
23. Tulloch J, Zamani R, Akrami M. Machine Learning in the Prevention, Diagnosis and Management of Diabetic Foot Ulcers: A Systematic Review. IEEE Access 2020;8:198977-199000. DOI: 10.1109/ACCESS.2020.3035327.
24. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015;350:g7647. DOI: 10.1136/bmj.g7647.
25. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10. DOI: 10.1186/1471-2288-7-10.
26. Dovell G, Staniszewska A, Ramirez J, et al. A systematic review of outcome reporting for interventions to treat people with diabetic foot ulceration. Diabet Med 2021;38(10):e14664. DOI: 10.1111/dme.14664.
27. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210. DOI: 10.1186/s13643-016-0384-4.
28. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. DOI: 10.1136/bmj.l4898.
29. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158(4):280-6. DOI: 10.7326/0003-4819-158-4-201302190-00009.
30. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 2019;170(1):W1-W33. DOI: 10.7326/M18-1377.
31. Stefanopoulos S, Qiu Q, Ren G, et al. A Machine Learning Model for Prediction of Amputation in Diabetics. J Diabetes Sci Technol 2022:19322968221142899. DOI: 10.1177/19322968221142899.
32. Margolis DJ, Mitra N, Malay DS, et al. Further evidence that wound size and duration are strong prognostic markers of diabetic foot ulcer healing. Wound Repair Regen 2022;30(4):487-490. DOI: 10.1111/wrr.13019.
33. Austin AM, Ramkumar N, Gladders B, et al. Using a cohort study of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling. BMC Med Res Methodol 2022;22(1):300. DOI: 10.1186/s12874-022-01774-8.
34. Kim RB, Gryak J, Mishra A, et al. Utilization of smartphone and tablet camera photographs to predict healing of diabetes-related foot ulcers. Comput Biol Med 2020;126:104042. DOI: 10.1016/j.compbiomed.2020.104042.

35. Jung K, Covington S, Sen CK, et al. Rapid identification of slow healing wounds. Wound Repair Regen 2016;24(1):181-8. DOI: 10.1111/wrr.12384.
36. Xie P, Li Y, Deng B, et al. An explainable machine learning model for predicting in-hospital amputation rate of patients with diabetic foot ulcer. Int Wound J 2022;19(4):910-918. DOI: 10.1111/iwj.13691.
37. Wang S, Xia C, Zheng Q, Wang A, Tan Q. Machine Learning Models for Predicting the Risk of Hard-to-Heal Diabetic Foot Ulcers in a Chinese Population. Diabetes Metab Syndr Obes 2022;15:3347-3359. DOI: 10.2147/DMSO.S383960.
38. Wang S, Wang J, Zhu MX, Tan Q. Machine learning for the prediction of minor amputation in University of Texas grade 3 diabetic foot ulcers. PLoS One 2022;17(12):e0278445. DOI: 10.1371/journal.pone.0278445.
39. Du C, Li Y, Xie P, et al. The amputation and mortality of inpatients with diabetic foot ulceration in the COVID-19 pandemic and postpandemic era: A machine learning study. Int Wound J 2022;19(6):1289-1297. DOI: 10.1111/iwj.13723.
40. Husers J, Hafer G, Heggemann J, Wiemeyer S, John SM, Hubner U. Predicting the amputation risk for patients with diabetic foot ulceration - a Bayesian decision support tool. BMC Med Inform Decis Mak 2020;20(1):200. DOI: 10.1186/s12911-020-01195-x.
41. Husers J, Hafer G, Heggemann J, Wiemeyer S, John SM, Hubner U. Development and Evaluation of a Bayesian Risk Stratification Method for Major Amputations in Patients with Diabetic Foot Ulcers. Stud Health Technol Inform 2022;289:212-215. DOI: 10.3233/SHTI210897.
42. Poradzka AA, Czupryniak L. The use of the artificial neural network for three-month prognosis in diabetic foot syndrome. J Diabetes Complications 2023;37(2):108392. DOI: 10.1016/j.jdiacomp.2022.108392.
43. Kasbekar PU, Goel P, Jadhav SP. A Decision Tree Analysis of Diabetic Foot Amputation Risk in Indian Patients. Front Endocrinol (Lausanne) 2017;8:25. DOI: 10.3389/fendo.2017.00025.
44. Huang J, Yang J, Qi H, Xu M, Xu X, Zhu Y. Prediction models for amputation after diabetic foot: systematic review and critical appraisal. Diabetology & Metabolic Syndrome 2024;16(1):126. DOI: 10.1186/s13098-024-01360-6.
45. Zhang Y, Cramb S, McPhail SM, et al. Factors Associated With Healing of Diabetes-Related Foot Ulcers: Observations From a Large Prospective Real-World Cohort. Diabetes Care 2021;44(7):e143-e145. DOI: 10.2337/dc20-3120.
46. Yotsu RR, Pham NM, Oe M, et al. Comparison of characteristics and healing course of diabetic foot ulcers by etiological classification: Neuropathic, ischemic, and neuro-ischemic type. Journal of Diabetes and its Complications 2014;28(4):528-535. DOI: https://doi.org/10.1016/j.jdiacomp.2014.03.013.
47. Shin JY, Roh S-G, Lee N-H, Yang K-M. Influence of Epidemiologic and Patient Behavior–Related Predictors on Amputation Rates in Diabetic Patients: Systematic Review and Meta-Analysis. The International Journal of Lower Extremity Wounds 2017;16(1):14-22. DOI: 10.1177/1534734617699318.
48. Luo Y, Liu C, Li C, Jin M, Pi L, Jin Z. The incidence of lower extremity amputation and its associated risk factors in patients with diabetic foot ulcers: A meta-analysis. Int Wound J 2024;21(7):e14931. (In eng). DOI: 10.1111/iwj.14931.
49. Breiman L. Random Forests. Machine Learning 2001;45(1):5-32. DOI: 10.1023/A:1010933404324.
50. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5(9):1315-6. (In eng). DOI: 10.1097/JTO.0b013e3181ec173d.

51. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 2015;25(4):932-9. (In eng). DOI: 10.1007/s00330-014-3487-0.
52. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology 2014;14(1):137. DOI: 10.1186/1471-2288-14-137.

FACULDADE DE **MEDICINA**