Automated assessment of diabetic foot ulcers: Estimation of PWAT categories using radiomic features and machine learning

Evaluación automatizada de úlceras del pie diabético: Estimación de categorías PWAT mediante características radiómicas y aprendizaje automático

Marcelo Márquez-Sandoval[1] https://orcid.org/0000-0001-5639-8888
Benjamín Morales Carvajal[2] https://orcid.org/0009-0004-3449-7278
Julio Sotelo Parraguez[3] https://orcid.org/0000-0002-0915-5215
Ana Aguilera[1, 2*] https://orcid.org/0000-0003-0726-1759

Recibido 19 de octubre de 2025, aceptado 29 de diciembre de 2025
Received: October 19, 2025 Accepted: December 29, 2025

ABSTRACT

The assessment of diabetic foot ulcers is essential for guiding therapeutic interventions and predicting outcomes such as wound healing or the need for surgical procedures. In this study, an automated pipeline for quantifying six categories of the Photographic Wound Assessment Tool (PWAT) was developed and validated using a dataset of 1,153 expert-annotated photographic images of diabetic foot ulcers (DFUs). All images were uniformly resized to 256 × 256 pixels, and ninety-three radiomic features were extracted via the PyRadiomics library. Support Vector Machine, Random Forest, and XGBoost classifiers were trained on 75% of the dataset, with the remaining 25% held out for validation. Class imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE), random oversampling, and undersampling strategies. The optimal performance was achieved with XGBoost combined with random oversampling, yielding an accuracy of 78% and a macro-average F1-score of 0.77. These results demonstrate the feasibility of a reproducible, quantitative, and fully automated system for standardizing the evaluation of diabetic foot ulcers.

Keywords: Diabetic foot ulcers (DFUs), photographic wound assessment tool (PWAT), pyradiomics, classification.

RESUMEN

La evaluación de las úlceras del pie diabético es esencial para guiar las intervenciones terapéuticas y predecir resultados como la cicatrización de heridas o la necesidad de procedimientos quirúrgicos. En este estudio, desarrollamos y validamos un proceso automatizado para cuantificar seis categorías de la Herramienta Fotográfica de Evaluación de Heridas (PWAT) utilizando un conjunto de datos de 1153 imágenes fotográficas de úlceras del pie diabético (UPD) anotadas por expertos. Todas las imágenes se redimensionaron uniformemente a 256 × 256 píxeles, y se extrajeron noventa y tres características

[1] Universidad de Valparaíso. Centro Interdisciplinario MEDING. Valparaíso, Chile. E-mail: marcelo.marquez@postgrado.uv.cl; ana.aguilera@uv.cl
[2] Universidad de Valparaíso, Escuela de Ingeniería Informática. Facultad de Ingeniería. Valparaíso, Chile. E-mail: benjamin.moralesc@estudiantes.uv.cl
[3] Universidad Técnica Federico Santa María. Departamento de Informática. Santiago, Chile. E-mail: julio.sotelo@usm.cl
* Autor de correspondencia: ana.aguilera@uv.cl

DOI: 10.64966/ingeniare.v33.24.

radiómicas mediante la biblioteca PyRadiomics. Los clasificadores de Máquina de Vectores de Soporte, Bosque Aleatorio y XGBoost se entrenaron en el 75% del conjunto de datos, y el 25% restante se reservó para validación. El desequilibrio de clases se abordó mediante la Técnica de Sobremuestreo Sintético de Minorías (SMOTE), sobremuestreo aleatorio y estrategias de submuestreo. El rendimiento óptimo se logró con XGBoost combinado con sobremuestreo aleatorio, lo que arrojó una precisión del 78% y un F1-score macropromedio de 0,77. Estos resultados demuestran la viabilidad de un sistema reproducible, cuantitativo y totalmente automatizado para estandarizar la evaluación de las úlceras del pie diabético.

Palabras clave: Úlceras del pie diabético (UPD), herramienta de evaluación fotográfica de heridas (PWAT), pyradiomics, clasificación.

INTRODUCTION

Effective management of diabetic foot ulcers (DFUs) is crucial for reducing associated complications. Diabetes currently affects more than 537 million people worldwide, and this number is projected to exceed 783 million by 2045 [1]. Approximately 25% of patients with diabetes will develop DFUs, with recurrence rates of 40% at one year and 65% at five years. In Chile, the prevalence of diabetes among adult patients is 12.3% [2].

DFUs represent the primary cause of non-traumatic amputations worldwide, resulting in significant clinical and economic burden [3]. Recent reports from the International Diabetes Federation indicate that up to 85% of amputations could be prevented through the implementation of standardized assessment and management protocols [4]. However, effective prevention depends on the use of objective and reproducible wound evaluation methods, as current manual assessments exhibit high clinical variability that directly impacts patient prognosis. This assessment allows clinicians to grade severity, guide treatment decisions, and predict outcomes, including healing and amputation risk [5]-[6]. Since 2010, several scoring instruments and prognostic classification systems have been refined to objectively evaluate chronic wounds in response to these challenges. The Photographic Wound Assessment Tool (PWAT) [5], which quantifies eight key wound attributes from photographic images, facilitating objective monitoring of wound healing progression [7]. Other validated instruments include the Bates-Jensen Wound Assessment Tool (BWAT) [8] and the TIME conceptual framework (Tissue, Infection/Inflammation, Moisture, Edge) [9], both of which are widely adopted to standardize chronic wound evaluation. Prognostic models such as SINBAD [10], WIfI (Wound, Ischemia, and foot Infection) [11], and the University of Texas classification [12]-[13] integrate multidimensional clinical parameters to generate individualized patient prognoses. This study presents the development and validation of an automated system for quantifying six of the eight PWAT items in DFUs (items 3-8), excluding area and depth measurements, which require dimensional calibration references (rulers and/or templates) typically unavailable in standard clinical photography datasets. The proposed framework integrates radiomic feature extraction, conventional machine learning models, and advanced class-balancing strategies to enhance diagnostic precision in patients with DFUs.

QUANTITATIVE WOUND ASSESSMENT PROTOCOLS

Six key approaches for the objective assessment of chronic wounds are discussed: the PWAT, BWAT, TIME framework, and the prognostic models SINBAD, WIfI, and the University of Texas classification. Table 1 presents the main characteristics of these approaches, comparing each method on diagnostic accuracy, ease of use, clinical applicability, and reproducibility. The PWAT is a digital instrument developed to evaluate wounds from photographic images [5]. This instrument includes eight classification items: wound area, depth, necrotic tissue type, necrotic tissue amount, exudate type, exudate amount, surrounding skin color, and wound edge characteristics. Each item is scored from 0 to 4, yielding a total score between 0 and 32. Higher scores indicate greater wound severity, while lower scores reflect improved healing. The BWAT provides a more comprehensive approach, scoring 13 parameters between 1 and 5 points, with total scores ranging from 13 to 65 points [8]. The

parameters assessed include size, depth, edges, necrosis, exudate, edema, induration, and wound-bed tissue characteristics. While this approach enables detailed severity profiling, it requires in-person assessment by trained clinicians, which limits scalability in clinical settings.

The TIME framework evaluates four key domains: Tissue viability, Infection/Inflammation, Moisture balance, and Edge advancement [9]. However, this approach depends on hands-on clinical examination, which restricts its applicability in remote settings [14]. Furthermore, the TIME mnemonic lacks a quantitative scoring component system and is therefore vulnerable to subjective interpretation. This limitation reduces its automation potential, although auxiliary assessment tools may improve clinical consistency. The prognostic models SINBAD, WIfI, and the University of Texas classification provide rapid risk stratification based on clinical indicators. However, these models depend on data that cannot be acquired exclusively from photographic images. SINBAD assigns binary scores to six items: site, ischemia, neuropathy, bacterial burden, area, and depth [10]. This system facilitates the prediction of non-healing or amputation outcomes with demonstrated clinical utility for 12-month healing assessment. Regardless, this system requires perfusion and neuropathy measurements, which are not discernible in photographs. This requirement limits the feasibility of image-based automation.

The WIfI system evaluates three components (Wound, Ischemia, and foot Infection) on a scale from 0 to 3 points to estimate amputation risk and inform vascular interventions [11]. The WifI system demonstrates strong correlations with clinical outcomes, with major amputation rates ranging from 5% to 50% and wound non-healing rates from 8% to 63% across clinical stages [15]. Nevertheless, the system's complexity and need for specialized instrumentation requirements limit its remote implementation. The University of Texas classification integrates grade and stage to characterize ulcer depth, infection, and ischemia [12]. Although widely used in clinical practice, its inter-observer reproducibility is moderate. Automating its application requires inferring systemic clinical data beyond image analysis capabilities.

RELATED WORK

In recent years, numerous studies have integrated wound assessment protocols with artificial intelligence (AI) techniques to enhance the diagnosis and monitoring of DFUs and other chronic lesions. Table 2 provides a comparative analysis of the most relevant studies in the literature, establishing

Table 1. Comparative overview of assessment and prognostic methods for DFUs.

Method	Diagnostic accuracy	Ease of use	Clinical applicability	Reproducibility
PWAT [5]	High intra-rater consistency.	Simple (8 items, photographic format).	Ideal; designed for remote assessments.	Good, due to clear visual criteria.
BWAT [8]	Provides a detailed profile; high sensitivity to identify specific issues.	Comprehensive but labor-intensive (13 items).	Limited; several items require in-person assessment.	Very high with proper training (agreement >90%).
TIME [9]	High qualitative sensitivity to detect barriers (e.g., devitalized tissue, infection).	Very easy to apply (mnemonic-based).	Moderate; partially applicable via images, but some aspects require direct contact.	Variable, depending on evaluator experience.
SINBAD [10]	Excellent for risk stratification; effectively predicts adverse outcomes.	Extremely simple (6 binary items).	High; most components observable remotely.	Moderate inter-observer, but robust in multicenter audits.
WIfI [11]	Highly accurate in predicting amputation and need for revascularization.	Complex; requires objective measurements (ABI, $TcPO_2$).	Limited; ischemia assessment is difficult remotely.	High in specialized teams, though data quality-dependent.
Texas [12]	Good prognostic capacity; improvement over Wagner scale.	Moderate; requires thorough physical examination (grade and stage).	Moderate; some elements are difficult to confirm via imaging.	Moderate, with variable inter-observer agreement without specific training.

benchmarks for the present contributions. Artificial Intelligence approaches applied to DFUs have predominantly focused on wound segmentation and area quantification as essential monitoring metrics. The 2022 Diabetic Foot Ulcer Segmentation Challenge (DFUSC 2022) demonstrated the current state-of-the-art in automated wound detection. The winning team achieved a Dice coefficient of 0.7287 for ulcer segmentation [16], highlighting the ongoing challenges in accurately delineating wound boundaries from clinical photographs. However, recent studies consistently report that the first two PWAT items, area and depth, are critical parameters. Their precise measurement requires in-situ calibration references, including graduated rulers, specialized templates, or advanced volumetric analysis technologies such as photogrammetry or point clouds.

Without appropriate reference points, the use of 2D images, even with approximate visual scales, yields inaccurate area and depth estimates, severely undermining clinical reliability. Lasschuit, Featherston, and Tonks [17], demonstrated that traditional two-dimensional measurement methods consistently underestimate depth changes in wounds. Although 3D cameras offer superior sensitivity and precision advantages, their high cost and technical complexity hinder their widespread clinical application. The study emphasizes that visual scales in 2D images do not fully address the inherent limitations of bidimensionality, particularly in wounds exhibiting curvature or variable depth. Similarly, multicenter trials such as the "Diabetic Foot Ulcer Photography Study" [18] demonstrated that any photograph-based assessment is fundamentally subjective and limited unless a properly positioned physical calibrator is included, particularly for complex or deep ulcers, even when processed under rigorous protocols.

Comparative analyses [19] indicate that, despite high statistical correlation between digital and three-dimensional methods, absolute accuracy, particularly in deep wounds, can be achieved only with techniques that provide factual spatial or volumetric information. These findings reinforce that image-only models, without metric or volumetric references, remain insufficient for robust, clinically integrable wound assessment of dimensional parameters. Multiple studies have investigated automated severity classification based on qualitative PWAT parameters to address these measurement limitations. Zhao *et al.* [6] leveraged items 3-8 of the PWAT to train classification models for tissue characteristics and wound appearance, achieving an accuracy of approximately 85% accuracy for individual attributes. Liu, Agu, Pedersen, Lindsay, Tulu, and Strong, [20] developed an end-to-end system for scoring morphological PWAT items using dense networks with attention mechanisms, accurately rating qualitative sub-attributes in approximately 80% of cases.

These studies face significant challenges due to class imbalance, as some severity grades are underrepresented in training datasets. Unlike protocol-based methods, recent detection approaches have emphasized binary classification tasks. Rathore, Kumar, Nandal, Dhaka, and Sharma [21], introduced a deep learning framework for DFUs identification, achieving 98.76% accuracy and 98.6% AUC using Siamese Neural Networks, while implementing explainability techniques (SHAP, LIME, Grad-CAM) to enhance clinical interpretability. Incorporating established prognostic scales into AI frameworks provides robust benchmarks for outcome prediction. Ha Van *et al.* [22] demonstrated that SINBAD score-based models achieve moderate predictive performance for non-healing outcomes. Conversely, Oliveira Cerqueira, Duarte, de Souza Barros, Cerqueira, and de Araújo [23], reported that higher WIfI stages significantly increase the risk of non-healing. These approaches require clinical data beyond image analysis, limiting their applicability for image-only assessment systems. The literature indicates that AI-based systems benefit significantly from standardized protocols, as consistent scales generate training labels, enabling reproducible mapping between image features and clinical assessments.

Nevertheless, several critical limitations persist: i) Training data dependency on manually-labeled datasets introduces intra- and inter-rater variability affecting model learning [14]; ii) Class imbalance leads to biased predictions toward prevalent categories [6]-[20]; iii) Deep learning models often function as "black boxes" with limited clinical interpretability [6]-[21]; iv) Large labeled datasets are required for adequate generalization, while medical datasets typically remain small and

institution specific. These challenges, together with the inherent limitations of 2D imaging for dimensional measurements, underscore the need to prioritize automated assessment on qualitative morphological features that can be reliably evaluated from photographic images.

In addition to these dataset and imaging constraints, the literature also emphasizes that segmentation performance must be interpreted using evaluation criteria aligned with the clinical objective and the type of error most relevant in practice. Müller, Soto-Rey, and Kramer [28], assert that metric selection must be aligned with clinically relevant error modes, such as boundary or region discrepancies, and advocate for transparent, standardized reporting to facilitate fair comparisons across studies. Likewise, Terven, Cordova-Esparza, Ramirez-Pedraza, Chavez-Urbiola, and J.A. Romero-Gonzalez [29] review commonly used loss functions and evaluation metrics in deep learning, highlight that no single metric is universally sufficient. They recommend combining complementary measures, such as overlap and boundary-based metrics, to achieve a more reliable assessment of segmentation performance. Collectively, these recommendations underscore the need for rigorous, multi-metric evaluation in DFU image analysis, particularly when models are intended to support clinically meaningful monitoring and decision-making.

MATERIALS

Dataset and annotation process

Images of diabetic foot ulcers (DFUs) were collected from three primary sources. Two experienced healthcare professionals filtered each dataset to exclude images that did not depict DFUs. The first dataset comprised 374 images from the Medetec repository [16], of which 179 were confirmed as DFUs. The second dataset comprised 1,180 images from the FUSeg2021 Challenge. The third dataset, from the Wuseg study [16], included 1,176 images representing various ulcer types such as diabetic, pressure, traumatic, venous, surgical, arterial, cellulitis, among other. Of these, 886 were verified to be DFUs.

Table 2. Comparative overview of studies on automated evaluation of chronic wounds.

Reference (Year)	Evaluation protocol	Applied AI technique	Performance metrics
Zhao *et al.* (2019) [6].	PWAT (depth and granulation categories).	Bilinear CNN for fine-grade classification across five levels (per item).	Accuracy 85% for depth; 85% for granulation (5 classes).
Shukla *et al.* (2020) [14].	BWAT (13 clinical items).	N/A (manual BWAT assessment).	BWAT score improvement: 13 vs. 5 points ($p < 0.05$) after oxygen therapy - used as a healing metric.
Sahdudak *et al.* (2025) [19].	SINBAD (0-6 scale).	N/A (prospective prognostic study).	One-year healing: 45% (SINBAD ≥ 3) vs. 78% (SINBAD < 3); AUC = 0.79.
Curti *et al.* (2024) [24].	PWAT (8 items, total score).	Segmentation + texture extraction + LASSO regression.	Spearman $\rho = 0.85$; MAE = 2.3 PWAT points.
Liu *et al.* (2024) [20].	PWAT (8 items, subscores).	Patch-based CNN with contextual attention	Accuracy = 80%; F1-score = 0.78.
Wang *et al.* (2022) [25].	IWGDF risk classification (classes 0 vs 1).	Bayesian classifier; SVM; Random Forest; MGWC.	Accuracy ≈ 87%; Sensitivity = 87%; Specificity = 87%.
Tao *et al.* (2025) [26].	Amputation risk prediction (multicenter clinical scores: UT, Wagner).	Logistic regression; Random Forest; XGBoost.	AUC = 0.90; Precision = 85%; Recall = 83%.
Khandakar *et al.* (2021) [27].	Thermogram-based DFU detection.	SVM; Random Forest; k-NN.	Accuracy = 90%; AUC = 0.88.
Rathore *et al.* (2025) [21].	DFU presence/absence detection.	DFU XAI: ensemble of six CNNs + explainability (SHAP, LIME, Grad-CAM).	Accuracy = 98.8%; Precision = 99.3%; Recall = 97.7%; F1-score = 98.5%; AUC = 98.6%.

After excluding non-DFUs cases, the final dataset consisted of 2,245 images (1,180 from FUSeg2021, 179 from Medetec, and 886 from Wuseg), each paired with its corresponding segmentation mask. A subset of 1,153 images was selected and labeled using the cloud-based data annotation platform Labelbox, which provides a collaborative, web-based environment for structured image annotation. Each image was independently annotated by two healthcare professionals with expertise in wound care. The annotation protocol, as illustrated in Figure 1, included bounding box segmentation of the lesion and scoring of PWAT items 3 through 8, following the criteria established in the Photographic Wound Assessment Tool. Items 1 and 2 (area and depth) were excluded due to technical limitations inherent to 2D imaging systems, as accurate quantification of these parameters require calibrated reference objects or volumetric analysis techniques [17]-[18]. Without in-situ calibration references, such as graduated rulers or standardized templates, estimating these parameters from photographic images introduces significant measurement uncertainty, compromising clinical reliability. Discrepancies between annotators were resolved through discussion and consensus to maintain annotation consistency. The final annotations were exported in *JSON* format and subsequently converted to CSV files for statistical analysis and the implementation of machine learning algorithms.

Characteristic matrix

The feature extraction process resulted in a structured dataset with dimensions of 1,153 × 7. Each row corresponds to an annotated image, and each column represents the filename identifier and six morphological assessment parameters (PWAT items 3-8). As shown in Figure 2, each feature variable was discretized into ordinal categories from 0 to 4. This

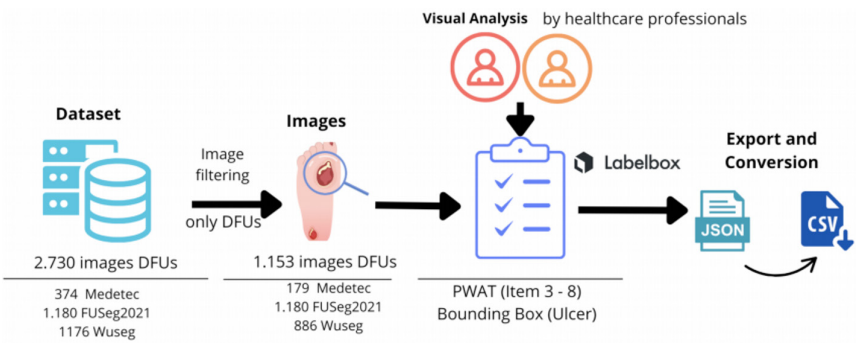

Figure 1. DFUs image using Labelbox. Bounding box and PWAT scoring interface displayed.

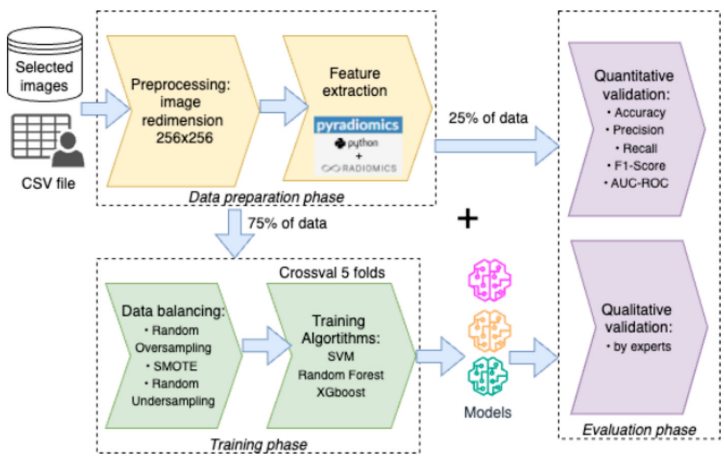

Figure 2. Pipeline for PWAT category prediction.

approach establishes a standardized encoding scheme for automated analysis and maintains the clinical significance of severity gradations. The following categorical definitions were implemented according to the established PWAT assessment framework [5]:

- **Item 3. Type of Necrotic Tissue:** Classifies the predominant devitalized tissue on the wound bed, ranging from no necrosis (score 0) to firm black eschar (score 4).
- **Item 4. Quantity of Necrotic Tissue:** Measures the percentage of the wound surface covered by necrotic or fibrinous tissue, from 0% (score 0) to 100% (score 4).
- **Item 5. Type of Granulation Tissue:** Describes the quality of reparative granulation tissue, from healthy, firm, red tissue (score 0) to pale, friable, or absent granulation (score 4).
- **Item 6. Quantity of Granulation Tissue:** Quantifies the proportion of the wound occupied by active granulation tissue, expressed as the percentage of wound-surface coverage (0% = score 0; 100% = score 4).
- **Item 7. Wound-Edge Condition:** Assesses the epithelial integrity at the wound margins –such as adherence, re-epithelialization, thickening, or undermining– to evaluate the potential for closure (score 0-4).
- **Item 8. Periulcer Skin Viability:** Identifies pathological changes in the surrounding skin (e.g., erythema, edema, maceration, dressing-related injury), scoring the number of signs present from 0 to 4.

METHODS

The approach for predicting PWAT category consists of three distinct phases: data preparation, training, and evaluation (Figure 2). During the data preparation phase, all images are resized and radiomic features are extracted from each one using PyRadiomics. The training phase uses 75% of the feature matrix and applies class-imbalance correction techniques, such as oversampling or undersampling to independently train three classifiers: Support Vector Machine (SVM), Random Forest, and XGBoost. Each classifier's performance is evaluated via five-fold cross-validation. In the final evaluation phase, the remaining 25% of the feature matrix serves as the test set. Model performance is validated using quantitative metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC, as well as blinded qualitative review by domain experts. The following sections provide detailed descriptions of the methods, algorithms, and metrics employed in this study.

Feature extraction

PyRadiomics [30] is an open-source platform for extracting a large panel of engineered features from medical images. Among the most widely used feature families available are the following:

- ***First-order features:*** *describe the global intensity distribution (mean, variance, percentiles, energy, entropy).*
- ***GLCM features:*** *capture texture homogeneity and contrast patterns using gray-level co-occurrence matrices.*
- ***GLRLM features:*** *quantify run lengths and uniformity of consecutive pixels sharing the same gray level.*
- ***GLSZM features:*** *assess the size and distribution of contiguous zones with identical intensity.*
- ***GLDM features:*** *describe gray-level dependence, i.e., the number of neighboring pixels required to achieve a specified intensity similarity.*
- ***NGTDM features:*** *capture the relationship of each pixel to the average intensity of its immediate neighborhood (e.g., coarseness, busyness, local contrast).*

Table 3. Frequency distribution of each score value by PWAT category in the feature matrix.

PWAT (Score)	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
0	110	118	28	32	8	8
1	192	300	158	164	300	357
2	440	380	414	451	535	447
3	281	194	401	321	253	255
4	120	151	142	175	46	76

Class-Balancing techniques

Class balancing techniques enable systematic evaluation of model performance across various sampling strategies, supporting the identification of configurations that significantly improve classification metrics for underrepresented classes, particularly F1-score and sensitivity [31]-[32]. The implemented techniques include *Random Oversampling*, which involves random duplication of minority class instances to achieve balanced class distributions [31]; *SMOTE (Synthetic Minority Over-sampling Technique)*, which generates synthetic samples through interpolation between existing minority class observations [32]; and *Random Undersampling*, which randomly removes majority class samples to equalize class distributions [31].

Machine-Learning algorithms

Three classical supervised classification algorithms were employed in this study:

- **Support Vector Machines (SVM):** selected for their effectiveness in high-dimensional spaces and robustness to slight data variations [33]. This method identifies the optimal separating hyperplane in high-dimensional feature spaces.
- **Random Forest (RF):** due to its ability to model complex nonlinear relationships and handle redundant features without extensive hyperparameter tuning [34]. This approach constructs an ensemble of decision trees aggregated by majority voting.
- **eXtreme Gradient Boosting (XGBoost):** an efficient tree-based method that iteratively optimizes residual errors, has been extensively validated for medical classification tasks and noisy or limited data scenarios [35].

Evaluation Metrics:

The following metrics were employed to assess training performance*:* Accuracy, which measures the proportion of correct predictions among all predictions; Precision (per class): the ratio of true positives to the total instances classified as positive by the model; Recall (per class): the ratio of true positives to the total actual positive instances; F1-Score (macro-average): the harmonic mean of precision and recall, providing a robust performance measure in imbalanced class settings; AUC-ROC (Area Under the Receiver Operating Characteristic Curve), which depicts the trade-off between true positive rate and false positive rate across different decision thresholds.

Validation framework

Two assessment schemes were utilized: *Data splitting* [36], which involves partitioning the dataset into separate training and evaluation subsets to estimate model performance on unseen data. *Cross-validation* [37]: is an evaluation technique in which the data are divided into k folds (commonly k = 5 or k = 10). The model is trained on k-1 folds and validated on the remaining fold, with this process repeated so that each iteration obtains a more robust estimate of average performance.

EXPERIMENTS

In our implementation, the training phase was conducted using Google Colab equipped with an NVIDIA A100 GPU, facilitating access to high-performance computing infrastructure. In terms of software, we have used an environment: Python 3.9.21 and its libraries for feature extraction (PyRadiomics 3.1.0, SimpleITK 2.3.1, pynrrd 1.0.0), classification training (scikit-learn 1.6.1, XGBoost 2.1.4), data balancing (imblearn 0.13.0), and others (numpy 1.26.4, pandas 2.2.1, matplotlib 3.8.3).

Data preparation phase

Following the recommendations of the Image Biomarker Standardization Initiative (IBSI) to ensure radiomic biomarker reproducibility [38], all original images were resized to 256 × 256 px to standardize spatial resolution. The image and corresponding masks were processed using *RadiomicsFeatureExtractor* from PyRadiomics, employing the default settings: *normalize*: false; *resampledPixelSpacing*: None; *interpolator*: sitkB-Spline; *binWidth set to* 25, among others.

All radiomic feature families were enabled, yielding a total of 93 features: first-order statistics (n = 18), gray-level co-occurrence matrix (GLCM; n = 24), gray-level run length matrix (GLRLM; n = 16), gray-level size zone matrix (GLSZM; n = 16), gray-level dependence matrix (GLDM; n = 14), and neighboring gray tone difference matrix (NGTDM; n = 5). These features were subsequently used for data balancing and classifier training.

Training phase

The 1,153 × 7 feature matrix was partitioned by category, yielding six separates *.csv* files. Each file contained two columns: the image identifier (file name) and the corresponding category score. This approach facilitated the training of dedicated classifiers for each clinical dimension of the PWAT. Two experiments were conducted: one without data balancing and one with data-balancing strategies. First, the three classifiers (SVM, RF, XGBoost) were trained and evaluated on each of the six unbalanced datasets, resulting in 18 baseline models. Subsequently, the three classifiers were combined with balancing techniques (Random Oversampling, SMOTE, Random Undersampling) for each of the six unbalanced datasets, yielding 54 additional models. Figure 3 presents the configuration employed in these data-balancing experiments.

Evaluation phase

The following results pertain to the evaluation phase of the proposed method.

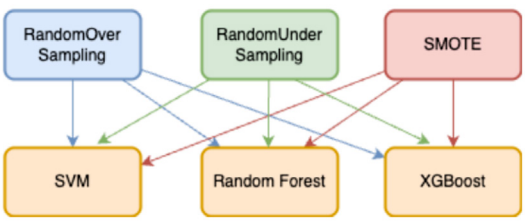

Figure 3. Diagram of the integration framework combining data-balancing techniques with classifiers.

Quantitative validation: Table 4 summarizes the accuracy obtained for the experiments conducted on the unbalanced datasets. Performance across all three classifiers, SVM, XGBoost, and RF, is largely comparable, with accuracy scores ranging from approximately 0.31. to 0.46, depending on the PWAT category. The highest accuracy occurs in Category 6, granulation tissue amount, where SVM attains 0.458, followed by RF at 0.430 and XGBoost at 0.406. Conversely, Category 4 necrotic tissue amount, proves to be most challenging, with accuracy values ranging 0.311-0.381. RF shows modestly higher accuracy in Categories 5, granulation tissue type, and 7, wound edge status, achieving 0.413 and 0.465, respectively. XGBoost leads slightly in Category 3, necrotic tissue type, with 0.409.

Due to the significant imbalance observed in the frequency distribution across categories in Table 3, data-balancing techniques were implemented. The primary objective was to balance the underrepresented classes, thereby enabling the models to better capture patterns in minority data and mitigate classification bias. Table 5 displays the number of samples per category after applying the balancing technique. Each entry corresponds to the count for scores 0 through 4.

Table 6 summarizes the performance metrics obtained after applying data-balancing techniques during model training. All models achieved substantial improvements in overall metrics compared to those trained on unbalanced datasets, where no model exceeded an accuracy of 0.41 due to the dominance

Table 4. Accuracy of the models on Original Data.

PWAT (Score)	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
SVM	0.402	0.381	0.353	0.458	0.430	0.409
XGBoost	0.409	0.311	0.360	0.406	0.416	0.392
Random Forest	0.406	0.371	0.413	0.430	0.465	0.392

Table 5. Results of the Data-Balancing method by Category.

Balancing Method	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Cat. 8
SMOTE	440	380	414	451	535	447
ROS	440	380	414	451	535	447
RUS	110	118	28	32	8	8

Table 6. Performance metrics of trained models.

Categories	Models	Algorithm	Recall	F1 Score	Accuracy	Precision	ROC AUC	Cross-Validation
Category 3	SVM	SMOTE	0.5182	0.5088	0.5182	0.5091	0.6980	0.4870
		Random Over Sampler	0.4673	0.4601	0.4673	0.4687	0.6679	0.4812
		Random Under Sampler	0.3623	0.3790	0.3623	0.4245	0.5975	0.2840
	XGBoost	SMOTE	0.6455	0.6383	0.6455	0.6373	0.7761	0.6440
		Random Over Sampler	0.7836	0.7771	0.7836	0.7757	0.8647	0.7352
		Random Under Sampler	0.3478	0.3661	0.3478	0.4114	0.5910	0.3278
	RandomForest	SMOTE	0.6218	0.6077	0.6218	0.6055	0.7616	0.6180
		Random Over Sampler	0.7655	0.7531	0.7655	0.7536	0.8535	0.7200
		Random Under Sampler	0.3986	0.4163	0.3986	0.4620	0.6240	0.3230
Category 4	SVM	SMOTE	0.4253	0.4145	0.4253	0.4198	0.6448	0.4456
		Random Over Sampler	0.4084	0.4156	0.4084	0.4336	0.6328	0.4196
		Random Under Sampler	0.3446	0.3394	0.3446	0.3369	0.5831	0.3212
	XGBoost	SMOTE	0.5705	0.5681	0.5705	0.5700	0.7350	0.5558
		Random Over Sampler	0.7095	0.7015	0.7095	0.6998	0.8202	0.6716
		Random Under Sampler	0.3108	0.3048	0.3108	0.3101	0.5666	0.3483
	RandomForest	SMOTE	0.5832	0.5771	0.5832	0.5775	0.7409	0.5551
		Random Over Sampler	0.7158	0.7098	0.7158	0.7096	0.8247	0.6519
		Random Under Sampler	0.3108	0.3142	0.3108	0.3246	0.5661	0.3574
Category 5	SVM	SMOTE	0.5135	0.4813	0.5135	0.4898	0.7023	0.5225
		Random Over Sampler	0.5135	0.4874	0.5135	0.4842	0.7022	0.4929
		Random Under Sampler	0.2286	0.2178	0.2286	0.2940	0.5405	0.2952
	XGBoost	SMOTE	0.6351	0.6233	0.6351	0.6210	0.7774	0.6289
		Random Over Sampler	0.7220	0.7053	0.7220	0.6995	0.8323	0.7023
		Random Under Sampler	0.2857	0.2988	0.2857	0.3392	0.5491	0.3333
	RandomForest	SMOTE	0.6371	0.6221	0.6371	0.6262	0.7789	0.6102
		Random Over Sampler	0.7336	0.7154	0.7336	0.7134	0.8406	0.6978
		Random Under Sampler	0.1714	0.1759	0.1714	0.2420	0.4986	0.3619
Category 6	SVM	SMOTE	0.5585	0.5273	0.5585	0.5316	0.7248	0.5328
		Random Over Sampler	0.5230	0.4939	0.5230	0.4932	0.7018	0.5257
		Random Under Sampler	0.2250	0.2394	0.2250	0.2738	0.5171	0.3167

Categories	Models	Algorithm	Recall	F1 Score	Accuracy	Precision	ROC AUC	Cross-Validation
Category 6	XGBoost	SMOTE	0.6631	0.6577	0.6631	0.6606	0.7892	0.6363
		Random Over Sampler	0.7801	0.7732	0.7801	0.7761	0.8649	0.7380
		Random Under Sampler	0.2500	0.2526	0.2500	0.2706	0.5292	0.3250
	RandomForest	SMOTE	0.6684	0.6548	0.6684	0.6519	0.7930	0.6221
		Random Over Sampler	0.7624	0.7532	0.7624	0.7556	0.8542	0.7357
		Random Under Sampler	0.2250	0.2242	0.2250	0.2475	0.5098	0.3417
Category 7	SVM	SMOTE	0.6697	0.6606	0.6697	0.6543	0.7904	0.6301
		Random Over Sampler	0.6323	0.6234	0.6323	0.6201	0.7667	0.5892
		Random Under Sampler	0.2000	0.2000	0.2000	0.2500	0.5000	0.2667
	XGBoost	SMOTE	0.7638	0.7604	0.7638	0.7593	0.8504	0.7228
		Random Over Sampler	0.8236	0.8235	0.8236	0.8237	0.8879	0.7906
		Random Under Sampler	0.1000	0.0800	0.1000	0.0667	0.4375	0.3667
	RandomForest	SMOTE	0.7578	0.7531	0.7578	0.7498	0.8463	0.7139
		Random Over Sampler	0.8371	0.8331	0.8371	0.8322	0.8960	0.7817
		Random Under Sampler	0.4000	0.3133	0.4000	0.2667	0.6250	0.2667
Category 8	SVM	SMOTE	0.6064	0.5952	0.6064	0.5920	0.7513	0.5805
		Random Over Sampler	0.5814	0.5725	0.5814	0.5691	0.7351	0.5609
		Random Under Sampler	0.3000	0.2905	0.3000	0.3400	0.5625	0.2667
	XGBoost	SMOTE	0.6762	0.6758	0.6762	0.6848	0.7967	0.6510
		Random Over Sampler	0.7639	0.7669	0.7639	0.7733	0.8495	0.7375
		Random Under Sampler	0.3000	0.2600	0.3000	0.2333	0.5625	0.3000
	RandomForest	SMOTE	0.6816	0.6746	0.6816	0.6727	0.7976	0.6390
		Random Over Sampler	0.7764	0.7783	0.7764	0.7838	0.8579	0.7172
		Random Under Sampler	0.2000	0.1800	0.2000	0.1667	0.5000	0.3667

of the majority classes. Figure 4 presents the row-normalized confusion matrices for the top-performing classifiers in each PWAT category after data balancing. The models exhibit consistent misclassification of classes 2 and 3 across all categories.

Qualitative validation: For the qualitative evaluation, three images not included in the training set were randomly selected, and a previously trained model was applied to generate their masks. The system's predictions were subsequently computed. These unseen validation images (Figure 5) were then compared with the automatically segmented masks and the estimated PWAT scores. In each case, the six clinical PWAT categories –necrotic tissue type and amount, granulation tissue type and amount, edge condition,

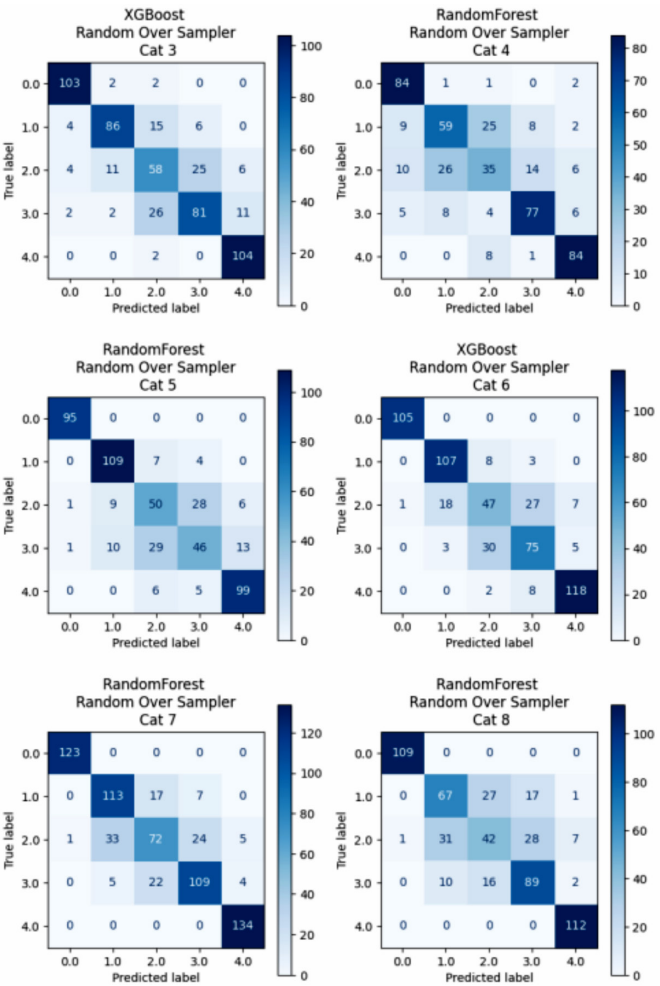

Figure 4. Confusion matrix across all categories (Models + augmented data).

Figure 5. Original photograph (left) and corresponding prediction mask (right) for the DFUs sample. The predicted PWAT scores for categories 3-8 were: 0, 2, 4, 4, 2, and 2, respectively.

and periulcer skin viability– were assessed, and the scores were recorded side-by-side. The maximum discrepancy between expert and model ratings did not exceed ± 1 point across all 18 observations. For example, the necrotic tissue amount score decreased from 3 to 2, whereas in Figure 6, the necrosis extent remained identical (4 vs. 4). Minor underestimation of necrotic regions in low-contrast areas and slight overestimation of granulation tissue along irregular borders Figure 5 were observed. However, these did not affect the overall assessment. The Spearman correlation coefficient for all paired scores (ρ = 0.935; p-value < 1×10^{7}) provides robust statistical evidence of the strong concordance between manual and automated ratings.

DISCUSSION

The proposed automated evaluation of Diabetic Foot Ulcers (DFUs) using a Pressure Wound Assessment Tool (PWAT)-based framework offers significant potential to standardize and objectively monitor chronic wounds by integrating machine learning techniques and data-balancing strategies. This approach is distinguished by its use of radiomics features and algorithms that require less memory.

Several methodological issues demand rigorous scrutiny to ensure accurate interpretation of the results. The PWAT category frequency distribution (Table 3) reveals a significant class imbalance: intermediate categories (scores 1 and 2) predominated the dataset, while extreme categories (0 and 4) were markedly underrepresented. This asymmetry initially biased the classifiers toward majority-class predictions, constraining their ability to detect wounds at more severe or advanced stages. Thus, a suite of data-balancing techniques –including Random Oversampling, SMOTE, and Random Undersampling– was implemented. Among these, SMOTE and Random Oversampling achieve the greatest improvements in addressing class imbalance (Table 4). For example, in the PWAT Category 8, the Random Forest algorithm combined with Random Oversampling achieved 78% accuracy with class balancing, compared to 39% accuracy without it. Moreover, Random Forest and XGBoost were the best-performing algorithms in PWAT classification, combined with the class-balancing method based on Random Oversampling.

The automated assessment was restricted to PWAT items 3-8 because visual annotations are more reliable for granulation tissue, necrotic tissue, edge condition, and perilesional skin quality in photographic images. Quantitative metrics, including wound area and absolute depth (items 1 and 2), were excluded because standardized protocols for their extraction from the available image repository were not established. Moreover, recent studies consistently indicate that the first two PWAT items, area and depth, are critical parameters that require precise

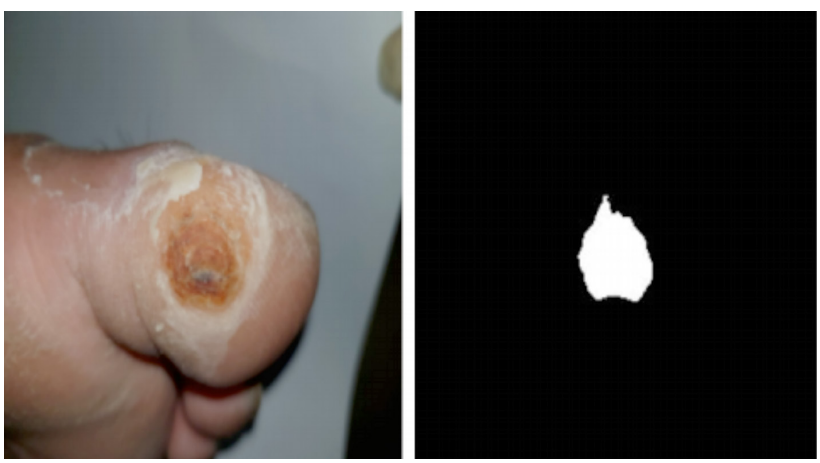

Figure 6. Original photograph (left) and corresponding prediction mask (right) for the DFUs sample. The predicted PWAT scores for categories 3-8 were: 0, 2, 4, 4, 2, and 2, respectively.

measurement using in-situ calibration references, such as graduated rulers, specialized templates, or advanced volumetric analysis technologies, such as photogrammetry and point clouds [24]-[39].

Within this framework, classical machine-learning models, such as Decision Trees, Random Forests, and XGBoost, offer distinct advantages over deep-learning "black-box" approaches. These methods inherently demand minimal computational resources and facilitate deployment on constrained hardware platforms. The rule-based architecture of these models enhances transparency, enabling clinicians to trace each decision pathway and offering an optimal balance between predictive performance and interpretability. For feature representation, PyRadiomics was employed to extract over 90 quantitative texture and intensity descriptors. The integration of robust data-balancing strategies, externally validated classical classifiers, and comprehensive radiomic feature extraction yields a practical and scalable method for automated PWAT estimation. These attributes position the proposed framework as a valuable adjunct to clinical workflows and highlight the imperative for continued validation across diverse patient populations prior to routine clinical adoption.

CONCLUSION

The integration of machine learning techniques with data-balancing strategies significantly enhances the automated classification of DFUs according to specific PWAT categories. Synthetic augmentation methods, particularly SMOTE and Random Oversampling, were essential for mitigating class-imbalance bias and resulted in substantial improvements in sensitivity and macro-average F1-score for critical minority classes. Focusing on selected PWAT items (categories 3 through 8) allowed emphasis on visually relevant aspects of wound healing. However, it is acknowledged as a limitation that dimensions related to wound size and absolute depth were not included. The issues excluded from this study represent potential areas for future research. In conclusion, the obtained results support the feasibility of automated image-based wound assessment systems as a complement to traditional clinical evaluation. The integration of these models into telemedicine platforms or specialized wound clinics has the potential to standardize assessments, streamline follow-up workflows, and improve therapeutic outcomes for patients with DFUs. Nevertheless, future multicenter and prospective cohort validations will be essential to ensure robustness, applicability, and clinical acceptance. Another promising direction involves employing automated and specialized methods to estimate the first two PWAT items, area and depth, as visual methods are insufficient for this task when labeling based on 2D images.

ACKNOWLEDGMENTS

A. Aguilera thanks to ANID, Chile, Regular Fondecyt No. Nº 1250344.
J. Sotelo thanks to ANID Millennium Science Initiative Program ICN2021 004, and the Department of Medical Imaging and Radiation Sciences at Monash University.
Marcelo Márquez thanks to FIBUV program at Universidad de Valparaíso.

REFERENCES

[1] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045", *Diabetes Research and Clinical Practice*, vol. 183, art. 109119, 2022, doi: 10.1016/j.diabres.2021.109119.

[2] M. Moreno González, "Situación de la diabetes mellitus y la obesidad en Chile", *Revista de la Sociedad Argentina de Diabetes*, vol. 56, no. 3, p. 38, 2022, doi: 10.47196/diab.v56i3sup.529.

[3] G. Reiber, B. Lipsky, and G. Gibbons, "The burden of diabetic foot ulcers," *The American Journal of Surgery*, vol. 176, no. 2A, pp. 5S-10S, 1998, doi: 10.1016/S0002-9610(98)00181-0.

[4] D. Armstrong, A. Boulton, and S. Bus, "Diabetic foot ulcers and their recurrence," *New England Journal of Medicine*, vol. 376, no. 24, pp. 2367-2375, 2017, doi: 10.1056/NEJMra1615439.

[5] N. Thompson, L. Gordey, H. Bowles, N. Parslow, and P. Houghton, "Reliability and validity of the revised photographic wound assessment tool on digital images taken of various types of chronic wounds," *Advances in Skin & Wound Care*, vol. 26,

no. 8, pp. 360-373, 2013, doi: 10.1097/01.ASW.0000431329.50869.6f.

[6] X. Zhao *et al.*, "Fine-Grained diabetic wound depth and granulation tissue amount assessment using bilinear convolutional neural network," *IEEE Access*, vol. 7, pp. 179151-179162, 2019, doi: 10.1109/ACCESS.2019.2959027.

[7] P. Houghton, C. Kincaid, K. Campbell, M. Woodbury, and D. Keast, "Photographic assessment of the appearance of chronic pressure and leg ulcers," *Ostomy Wound Management*, vol. 46, no. 4, pp. 20-26, 28-30, 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10788924/

[8] B. Bates-Jensen, H. McCreath, D. Harputlu, and A. Patlan, "Reliability of the Bates-Jensen wound assessment tool for pressure injury assessment: The pressure ulcer detection study," *Wound Repair and Regeneration*, vol. 27, no. 4, pp. 386-395, 2019, doi: 10.1111/wrr.12714.

[9] R. Harries, D. Bosanquet, and K. Harding, "Wound bed preparation: TIME for an update," *International Wound Journal*, vol. 13, no. S3, pp. 8-14, Apr., 2016, doi: 10.1111/iwj.12662.

[10] P. Ince, Z. Abbas, J. Lutale, A. Basit, S.M. Ali, F. Chohan, S. Morbach, J. Möllenberg, F. Game, and W. Jeffcoate, "Use of the SINBAD classification system and score in comparing outcome of foot ulcer management on three continents," *Diabetes Care*, vol. 31, no. 5, pp. 964-967, 2008, doi: 10.2337/dc07-2367.

[11] J. Mills, M. Conte, D.G. Armstrong, F. Pomposelli, A. Schanzer, A. Sidawy, and G. Andros, "The society for vascular surgery lower extremity threatened limb classification system: Risk stratification based on wound, ischemia, and foot Infection (WIfI)," *Journal of Vascular Surgery*, vol. 59, no. 1, pp. 220-234.e2, 2014, doi: 10.1016/j.jvs.2013.08.003.

[12] D. Armstrong, L. Lavery, and L. Harkless, "Validation of a diabetic wound classification system: The contribution of depth, infection, and ischemia to risk of amputation," *Diabetes Care*, vol. 21, no. 5, pp. 855-859, 1998, doi: 10.2337/diacare.21.5.855.

[13] T. Santema, E. Lenselink, R. Balm, and D. Ubbink, "Comparing the Meggitt-Wagner and the University of Texas wound classification systems for diabetic foot ulcers: inter-observer analyses," *International Wound Journal*, vol. 13, no. 6, pp. 1137-1141, 2016, doi: 10.1111/iwj.12429.

[14] U. Shukla, A. Kumar, S. Anushapreethi, and S. Singh, "Evaluation of the efficacy of hyperbaric oxygen therapy in the management of diabetic ulcer using Bates-Jensen wound assessment tool," *Anesthesia: Essays and Researches*, vol. 14, no. 2, pp. 335-342, 2020, doi: 10.4103/aer.aer_68_20.

[15] D. Cull, G. Manos, M. Hartley, S. Taylor, E. Langan, J. Eidt, and B. Johnson, "An early validation of the society for vascular surgery lower extremity threatened limb classification system," *Journal of Vascular Surgery*, vol. 60, no. 6, pp. 1535-1542, Dec., 2014, doi: 10.1016/j.jvs.2014.08.107.

[16] R. Brüngel, C. Kendrick, B. Cassidy, B. Bracke, C. Friedrich, N. Reeves, J. Pappachan, and M. H. Yap, "Diabetic foot ulcers grand challenge, 4th challenge, DFUC 2024," in *Diabetic Foot Ulcers Grand Challenge, DFUC 2024, Proceedings*, Lecture Notes in Computer Science, M.H. Yap, C. Kendrick, and R. Brüngel, Eds., vol. 15335, 2025, pp. 109-124, doi: 10.1007/978-3-031-80871-5_10.

[17] J. Lasschuit, J. Featherston, and K. Tonks, "Reliability of a three-dimensional wound camera and correlation with routine ruler measurement in diabetes-related foot ulceration," *Journal of Diabetes Science and Technology*, vol. 15, no. 6, pp. 1361-1367, 2021, doi: 10.1177/1932296820974654.

[18] S. Brown *et al.*, "Diabetic foot ulcer photography study: a study within a trial to assess the reliability of two-dimensional (2D) photography for the assessment of ulcer healing in patients with diabetes-related foot ulcers-protocol paper," *BMJ Open*, vol. 15, no. 1, Art. no. e090299, 2025, doi: 10.1136/bmjopen-2024-090299.

[19] G. Sahdudak and U. Gunes, "Comparing digital, mobile and three-dimensional methods in pressure injury measurement: agreement in surface area and depth assessments," *Journal of Clinical Nursing*, vol. 35, no. 1, 2025, doi: 10.1111/jocn.17813.

[20] Z. Liu, E. Agu, P. Pedersen, C. Lindsay, B. Tulu, and D. Strong, "Chronic wound image

augmentation and assessment using semi-supervised progressive Multi-Granularity EfficientNet," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 5, pp. 404-420, 2024, doi: 10.1109/ojemb.2023.3248307.

[21] P. Rathore, A. Kumar, A. Nandal, A. Dhaka, and A.K. Sharma, "A feature explainability-based deep learning technique for diabetic foot ulcer identification," *Scientific Reports*, vol. 15, no. 1, Art. no. 6758, 2025, doi: 10.1038/s41598-025-90780-z.

[22] G. Ha Van *et al.*, "Use of the SINBAD score as a predicting tool for major adverse foot events in patients with diabetic foot ulcer: A French multicentre study," *Diabetes/Metabolism Research and Reviews*, vol. 39, no. 8, Art. no. e3705, 2023, doi: 10.1002/dmrr.3705.

[23] L. de Oliveira Cerqueira, E. Duarte, A. de Souza Barros, J.R. Cerqueira, and W.J.B. de Araújo, "Classificação WifI: o novo sistema de classificação da society for vascular surgery para membros inferiores ameaçados, uma revisão de literatura," *Jornal Vascular Brasileiro*, vol. 19, p. e20190070, 2020, doi: 10.1590/1677-5449.190070.

[24] N. Curti *et al.*, "Automated prediction of photographic wound assessment tool in chronic wound images," *Journal of Medical Systems*, vol. 48, no. 1, Art, no. 14, 2024, doi: 10.1007/s10916-023-02029-9.

[25] S. Wang, J. Wang, M. Zhu, and Q. Tan, "Machine learning for the prediction of minor amputation in University of Texas grade 3 diabetic foot ulcers," *PLOS ONE*, vol. 17, no. 12, Art. no. e0278445, 2022, doi: 10.1371/journal.pone.0278445.

[26] H. Tao, L. You, Y. Huang, Y. Chen, L. Yan, D. Liu, S. Xiao, B. Yuan, and M. Ren, "An interpreting machine learning models to predict amputation risk in patients with diabetic foot ulcers: a multi-center study," *Frontiers in Endocrinology*, vol. 16, Art. no. 1526098, 2025, doi: 10.3389/fendo.2025.1526098.

[27] A. Khandakar *et al.*, "A machine learning model for early detection of diabetic foot using thermogram images," *Computers in Biology and Medicine*, vol. 137, Art. no. 104838, Sep., 2021, doi: 10.1016/j.compbiomed.2021.104838.

[28] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," 2022, arXiv: 2202.05273.

[29] J. Terven, D.M. Cordova-Esparza, A. Ramirez-Pedraza, E.A. Chavez-Urbiola, and J.A. Romero-Gonzalez, "Loss functions and metrics in deep learning," 2023, arXiv: 2307.02694.

[30] J. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, Art. no. e104-e107, 2017, doi: 10.1158/0008-5472.can-17-0339.

[31] H. He and Y. Ma, *Imbalanced Learning*, Hoboken, NJ, USA: Wiley-IEEE Press, 2013, doi: 10.1002/9781118646106.

[32] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[33] A. Mammone, M. Turchi, and N. Cristianini, "Support vector machines," *WIREs Computational Statistics*, vol. 1, no. 3, pp. 283-289, 2009, doi: 10.1002/wics.49.

[34] T.K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, vol. 1, 1995, pp. 278-282, doi: 10.1109/icdar.1995.598994.

[35] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[36] A. Géron, "Create a Test Set," in *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. N. Tache, Ed., Sebastopol, CA, USA: O'Reilly Media, 2019, pp. 54-56.

[37] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40-79, 2010, doi: 10.1214/09-ss054.

[38] A. Zwanenburg *et al.*, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology*, vol. 295, no. 2, pp. 328-338, 2020, doi: 10.1148/radiol.2020191145.

[39] L. Jørgensen, U. Halekoh, G. Jemec, J. Sørensen, and K. Yderstræde, "Monitoring wound healing of diabetic foot ulcers using two-dimensional and three-dimensional wound measurement techniques: A prospective cohort study," *Advances in Wound Care*, vol. 9, no. 10, pp. 553-563, 2020, doi: 10.1089/wound.2019.1000.