**ETMHRJ** (2026) 1: 16-27
https://doi.org/.....

**Original Article**

# Explainable Artificial Intelligence (XAI) for Amputation Risk Prediction in Diabetic Foot Ulcers: A SHAP-based Interpretable Machine Learning Framework

**Amir Arshia Beheshti [1], Shafagh Ali Asgarzadeh [2*], Sevda Mikaeili Mirak [3], Vahideh Aghamohammadi [4]**

[1] *Students Research Committee, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran*
[2] *Department of Internal Medicine, School of Medicine, Ardabil University of Medical Sciences, Ardabil, Iran*
[3] *Department of Clinical Pharmacy, School of Pharmacy, Ardabil University of Medical Sciences, Ardabil, Iran*
[4] *Khalkhal University of Medical Sciences, Khalkhal, Iran*

## Abstract

**Introduction:** Diabetic foot ulcers (DFUs) are a major complication of diabetes, with substantial risk of amputation. Although the Wagner grading system is widely used for risk stratification, it has moderate inter-clinician agreement and does not incorporate systemic inflammatory biomarkers. Machine learning (ML) may improve prediction, but limited explainability and reliance on non-routine diagnostics restrict clinical adoption. This study evaluated explainable ML models using routinely available clinical variables in a Middle Eastern population.

**Methods:** This retrospective cohort study included 734 DFU patients treated at Ardabil University of Medical Sciences between 2017 and 2023, including 298 amputations (40.6%). Twelve routinely available clinical variables were selected using univariate analysis, correlation assessment, and variance inflation factor analysis. Ten ML algorithms were trained using an 80:20 stratified split, SMOTE, 10-fold cross-validation, and hyperparameter tuning. Models were evaluated using AUC-ROC, F1-score, calibration measures, and SHAP-based interpretability.

**Results:** CatBoost achieved the highest cross-validation AUC (0.973±0.021), followed by Stacking Ensemble (0.972±0.020) and XGBoost (0.969±0.020). On the independent test set, SVM showed the best discrimination (AUC=0.987, F1=0.920), followed by Random Forest (AUC=0.986) and Logistic Regression (AUC=0.982). The Wagner Grade-only model achieved AUC=0.975, indicating only modest incremental benefit from multivariate ML models. SHAP analysis identified Wagner Grade as the dominant predictor, accounting for approximately 85% of model importance. Nine models demonstrated good calibration.

**Conclusion:** Explainable ML models based on routine clinical variables accurately predicted amputation risk (AUC>0.97), but provided only modest improvement over Wagner Grade alone. External multicenter validation is required before clinical implementation.

**Keywords**: Diabetic foot ulcer; Amputation prediction; Machine learning

## Introduction

Diabetic foot ulcers (DFUs) affect an estimated 6.3% of people with diabetes worldwide [1] and are a major cause of lower-limb amputations, with an overall amputation rate of approximately 31% among affected patients [2]. With the global prevalence of diabetes projected to reach 10.9% (approximately 700 million adults) by 2045, diabetic foot ulcers are expected to impose a substantial public health burden [3]. In Iran and similar low- and middle-income countries, around 2-4% of diabetic patients develop DFUs each year, and

* Corresponding author. E-mail: sh.aliasghrzadeh@arums.ac.ir (S. Ali Asgarzadeh).
Received 10 June 2026; Revised 19 August 2026; Accepted 22 August 2026.

15-20% of them undergo amputation within 12 months [4,5]. Five-year mortality post-ulceration reaches 40-55%, exceeding that of many common cancers [6,7]. Annual direct healthcare costs in Iran alone exceed $1.8 billion [8].

The Wagner classification system remains the most widely used grading tool for assessing ulcer severity and informing treatment decisions. Wagner Grade ranges from 0 (intact skin) to 5 (gangrene of the entire foot), with higher grades generally associated with worse outcomes [9] (Figure1).



**Figure 1.** Wagner Grade ranges from 0 (intact skin) to 5 (gangrene of the entire foot).

However, Wagner grading has notable limitations: moderate inter-rater reliability (κ=0.60-0.70) [10], exclusion of systemic inflammatory markers (e.g., WBC, CRP, ESR), and a lack of individualized probabilistic risk estimates.

Machine learning (ML) models, particularly ensemble methods like XGBoost, CatBoost, and Random Forest, have demonstrated superior performance in predicting diabetic complications compared to traditional risk scores. Recent meta-analyses reported pooled AUCs exceeding 0.85 for ML-based DFU prediction. However, clinical uptake remains limited due to "black-box" opacity, dependence on advanced diagnostics (transcutaneous oxygen pressure [TcPO2], infrared thermography, microbiology cultures), and a lack of prospective validation in resource-limited settings [11].

Explainable artificial intelligence (XAI), particularly SHapley Additive exPlanations (SHAP) [12], offers a solution by quantifying individual feature contributions to model predictions. SHAP provides both global importance rankings and local, patient-level explanations, making ML outputs clinically interpretable. Despite this, few studies have applied XAI to DFU amputation prediction using only routine clinical variables in Middle Eastern populations, and none have systematically addressed data leakage while establishing Wagner Grade baselines for comparison.

This study aims to develop and validate explainable ML models for predicting amputation risk in DFU patients using exclusively routine clinical variables available in resource-limited settings. Data leakage concerns were systematically addressed by excluding outcome-related predictors. A Wagner Grade-only baseline was established to quantify incremental ML value. Rigorous feature selection was employed to minimize multicollinearity, and SHAP analysis was leveraged to identify clinically actionable risk factors.

## Materials & Methods

### Study Design and Setting

This retrospective cohort study analyzed electronic health records of adult patients diagnosed with diabetic foot ulcers at Ardabil University of Medical Sciences (Ardabil, Iran) between January 2017 and December 2023. The study received ethics approval from the institutional review board (IR. ARUMS. REC. 1402.123), with informed consent waived due to the retrospective use of de-identified data. Reporting adheres to TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) and CONSORT-AI (Consolidated Standards of Reporting Trials—Artificial Intelligence) guidelines.

### Patient Selection and Data Collection

Initial Cohort: Patients were identified via ICD-10 codes E11.621 (Type 2 diabetes with foot ulcer), E10.621 (Type 1 diabetes with foot ulcer), and L97 (chronic ulcer of lower limb). The initial cohort comprised 1,247 patients.

Exclusion Criteria (Sequential): Patients were excluded if they had:

1. Missing demographic data (age, sex): 183 patients excluded
2. Undocumented Wagner Grade classification: 156 patients excluded
3. Incomplete laboratory values (>20% missing): 98 patients excluded

4. Undocumented amputation outcomes: 76 patients excluded

Final Cohort: 734 patients with complete data underwent analysis (Supplementary Figure S1).

Data leakage prevention: The dataset was split into training (80%, n= 587) and test (20%, n= 147) sets using stratified random sampling (random state=42) BEFORE any feature selection or preprocessing. The 'OUTCOME' clinical score—a composite variable incorporating prognostic information overlapping with the amputation endpoint—was excluded from the feature set to prevent data leakage.

**Outcome Definition**

The primary outcome was major amputation, defined as any lower extremity amputation above the ankle (transtibial, transfemoral, or hip disarticulation) performed during the index hospitalization or within 12 months of initial ulcer diagnosis. Minor amputations (digital or transmetatarsal) were not classified as primary outcome events.

**Feature Selection and Preprocessing**

Initial Features: 23 candidate features were assessed, encompassing demographics (age, sex), clinical characteristics (Wagner Grade, diabetes duration, heart disease, kidney disease, hospitalization length), and laboratory parameters (WBC, hemoglobin, neutrophil, lymphocyte, monocyte, eosinophil, platelet, hematocrit, ESR, CRP, creatinine, urea, potassium, fasting blood sugar, RBC).

**Three-Stage Selection Pipeline**

Stage 1 - Univariate Testing: Mann-Whitney U tests (continuous variables) and chi-square tests (categorical variables) identified features with Bonferroni-adjusted $p<0.1$. Wagner Grade demonstrated the strongest association (test statistic=686.17, $p<10^{-90}$). This stage identified 19 significant features.

Stage 2 - Correlation Analysis: Features with Pearson $|r|>0.7$ were screened, removing those with lower univariate scores. This eliminated 4 redundant hematological markers: hemoglobin, neutrophil count, lymphocyte count, and hematocrit.

Stage 3 - Multicollinearity Assessment: Variance inflation factor (VIF) was computed iteratively on the training set only. Features with VIF>10 were removed: urea, potassium, and fasting blood sugar.

Final Feature Set: 12 independent predictors with acceptable multicollinearity (all VIF<8.0): WBC, heart disease, hemoglobin (HB), kidney disease, Wagner Grade, monocyte count, length of hospitalization, platelet count, creatinine, RBC, eosinophil count, and ESR.

Events-per-variable ratio: 24.83 (298 amputations / 12 features), exceeding the recommended minimum of 10-20 for stable logistic regression.

**Wagner Grade Baseline Model**

To quantify the incremental value of multivariate machine learning, a Wagner Grade-only baseline was established using logistic regression with Wagner Grade as the sole predictor. This represents current clinical practice, where amputation risk is primarily stratified by ulcer severity classification. All machine learning models were compared against this baseline using DeLong's test for AUC comparison.

**Machine Learning Models**

Ten algorithms were evaluated to capture diverse modeling approaches:

Tree-Based Ensembles:

XGBoost: Gradient-boosted trees with L1/L2 regularization

CatBoost: Gradient boosting with ordered target encoding for categorical features

LightGBM: Histogram-based gradient boosting with leaf-wise tree growth

Random Forest: Bagging ensemble of decision trees with bootstrap aggregation

AdaBoost: Sequential boosting with adaptive weighting of misclassified instances

Decision Tree: Single CART-based decision tree with Gini impurity

Linear/Kernel Methods:

Logistic Regression: L2-penalized logistic regression with saga solver

Support Vector Machine (SVM): Radial basis function (RBF) kernel with probability estimates

Instance-Based:
K-Nearest Neighbors (KNN): Distance-weighted voting with Euclidean metric
Meta-Ensemble:
Stacking Ensemble: Combined predictions from CatBoost, XGBoost, and LightGBM with Logistic Regression as meta-learner, trained via 10-fold cross-validation

**Hyperparameter Optimization**

RandomizedSearchCV with 5-fold stratified cross-validation (n_iter=20, scoring='roc_auc') optimized hyperparameters for each algorithm. Representative configurations: XGBoost: n_estimators=300, max_depth =5, learning_rate=0.05, subsample=1.0, colsampl e_ bytree=0.8, CatBoost: iterations=300, depth=5, learning_rate =0.05, l2_leaf_reg=3, LightGBM:n_estimators=200, max_depth=5, learning_rate=0.05, num_ leaves = 50, subsample=0.8 • SVM: C=10, gamma='sc ale', kernel='rbf', probability=True

**Class Imbalance Handling**

Synthetic Minority Oversampling Technique (SMOTE) was applied within cross-validation folds using imbalanced-learn 0.14.0 to address the 40.6% amputation prevalence. SMOTE parameters: k_neighbors=5, random_state=42. Scale-sensitive models (SVM, KNN, Logistic Regression) were preceded by RobustScaler to mitigate outlier influence.
Scale-sensitive models (SVM, KNN, Logistic Regression) were preceded by RobustScaler to mitigate outlier influence.
Discrimination Metrics:
• Area Under the Receiver Operating Characteristic Curve (AUC-ROC) with DeLong 95% confidence intervals
• F1-score (harmonic mean of precision and recall)
• Sensitivity, specificity, positive predictive value, negative predictive value
• Confusion matrices visualizing true/false positives/negatives
Statistical Comparison: DeLong's test with Bonferroni correction (α=0.05/55 pairwise comparisons) assessed statistical significance of AUC differences between models and the Wagner Grade baseline.
Calibration Assessment:
• Brier scores (lower indicates better calibration)
• Hosmer-Lemeshow goodness-of-fit tests (n_bins=10, quantile strategy; p>0.05 indicates good calibration)
• Calibration plots comparing predicted probabilities to observed amputation rates across deciles
Clinical Utility:
• Decision curve analysis (DCA) over threshold probabilities 0.05-0.60 to assess net benefit
• Youden's J statistic (sensitivity + specificity - 1) to determine optimal classification thresholds
Reproducibility: All analyses used random state=42 for reproducibility. Ten-fold stratified cross-validation ensured balanced class distribution across folds.

**Explainability Analysis**

SHAP (SHapley Additive exPlanations) values quantified feature contributions to model predictions, providing both global importance rankings and local, patient-level explanations.
Implementation:
• TreeExplainer for tree-based models (exact Shapley values)
• KernelExplainer for linear/kernel models (approximation with 100 background instances via k-means clustering)
• SHAP values computed on training set for global importance
• 50 randomly sampled test instances for local explanations
Visualizations:
• Beeswarm plots: Feature distributions colored by value (red=high, blue=low), x-axis showing SHAP value (impact on prediction)
• Bar plots: Mean absolute SHAP values ranked by global importance
• Dependence plots: Relationship between feature values and SHAP values, with interaction effects

**Statistical Analysis**

Baseline characteristics were compared between amputation and non-amputation groups using Mann-Whitney U tests for continuous variables and chi-square or Fisher's exact tests for categorical variables. Bonferroni correction adjusted for multiple comparisons. Shapiro-Wilk tests assessed normality. Multivariable logistic regression calculated adjusted odds ratios with 95% confidence intervals, visualized via forest plots. Significance was set at $\alpha$=0.05 (two-tailed).

**Software and Packages**

Analyses were conducted in Python 3.12.12 using scikit-learn 1.6.1, XGBoost 3.1.2, LightGBM 4.6.0, CatBoost 1.2.8, imbalanced-learn 0.14.0, SHAP 0.50.0, scipy 1.16.3, statsmodels 0.14.5, matplotlib 3.10.0, and seaborn 0.13.2.

**Results**

The final dataset comprised 734 diabetic foot ulcer patients (from 1,247 initially screened; 41.1% exclusion rate documented in Supplementary Figure S1) with 40.6% amputation prevalence (298/734). Stratified 80:20 splitting yielded 587 training and 147 test samples with balanced prevalence (40.5% vs. 40.8%, respectively). No missing values were detected in the final cohort, enabling complete-case analysis. Events-per-variable ratio was 24.83 (298 amputations / 12 features), exceeding the recommended minimum of 10-20 for stable logistic regression.

The three-stage selection pipeline reduced 23 candidate features to 12 final predictors. Stage 1 (univariate testing) identified 19 features with Bonferroni-adjusted $p<0.1$, led by Wagner Grade (test statistic=686.17, $p<10^{-90}$). Stage 2 (correlation analysis, $|r|>0.7$) eliminated 4 redundant hematological markers (hemoglobin, neutrophil, lymphocyte, hematocrit). Stage 3 (VIF>10 threshold) removed 3 metabolic variables (urea, potassium, fasting blood sugar). The "OUTCOME" variable was excluded to prevent data leakage. Final features showed acceptable multicollinearity (all VIF<8.0).

Between-group comparisons revealed significant differences in clinical and laboratory parameters. Wagner Grade showed the strongest association with amputation, with amputation patients exhibiting markedly higher grades (mean 3.45±1.54). Significant predictors also included WBC (mean 10071.25±4496.27, $p<0.0001$), heart disease (65% vs 35%, $p<0.0001$), length of hospitalization (mean 10.13±11.77 days, $p<0.0001$), and various inflammatory and hematological markers. All continuous variables demonstrated non-normal distributions (Shapiro-Wilk tests, $p<0.05$), justifying non-parametric statistical methods.

Ten-fold stratified cross-validation evaluated ten machine learning algorithms (Table 1). Tree-based methods achieved superior performance: • CatBoost: AUC=0.973±0.021, F1=0.895±0.054 • Stacking Ensemble: AUC=0.972±0.020, F1=0.894±0.070 XGBoost: AUC=0.969±0.020, F1=0.883±0.075 LightGBM: AUC=0.968±0.021, F1=0.875±0.089 RandomForest: AUC=0.966±0.028, F1=0.897±0.043 Simpler models showed moderate performance: SVM: AUC=0.963±0.026, F1=0.884±0.058 LogisticRegression:AUC=0.959±0.027, F1=0.842± 0.092. KNN demonstrated the weakest performance (AUC =0.890±0.036). All models used 10-fold stratified cross-validation with SMOTE. Models ranked by mean AUC.

**Table 1.** Ten-Fold Cross-Validation Performance Metrics

| Model | Mean | | Rank |
|---|---|---|---|
| | AUC±SD | F1±SD | |
| CatBoost | 0.973±0.021 | 0.895±0.054 | 1 |
| Stacking Ensemble | 0.972±0.020 | 0.894±0.070 | 2 |
| XGBoost | 0.969±0.020 | 0.883±0.075 | 3 |
| LightGBM | 0.968±0.021 | 0.875±0.089 | 4 |
| Random Forest | 0.966±0.028 | 0.897±0.043 | 5 |
| AdaBoost | 0.965±0.025 | 0.887±0.061 | 6 |
| SVM | 0.963±0.026 | 0.884±0.058 | 7 |
| Logistic Regression | 0.959±0.027 | 0.842±0.092 | 8 |
| Decision Tree | 0.955±0.041 | 0.895±0.065 | 9 |
| KNN | 0.890±0.036 | 0.768±0.084 | 10 |

On the held-out test set (n=147), models demonstrated strong generalization: SVM: AUC=0.987 (95% CI: 0.959-1.000), F1=0.920, Precision=0.981, Recall=0.867, RandomForest: AUC=0.986, F1=0.914, Precision=0.946, Recall=0.883, CatBoost: AUC= 0.984, F1=0.914, Precision=0.946, Recall=0.883,

LogisticRegression: AUC=0.982, F1= 0.913, Precision=0.866, Recall=0.967, Stacking Ensemble: AUC =0.982, F1=0.924, Precision=0.932, Recall=0.917 XGBoost: AUC=0.980, F1=0.906, Precision=0.930, Recall=0.883 (Table 2).

Cross-validation to test set AUC discrepancies were minimal (<0.02 for top models), confirming robust generalization.

The Wagner Grade-only logistic regression baseline achieved:

- Cross-validation AUC: 0.969±0.024
- Test set AUC: 0.975 (95% CI: 0.935-1.000)
- Sensitivity: 0.867, Specificity: 0.989

The best-performing multivariate model (SVM, test AUC=0.987) provided an incremental benefit of 1.22 percentage points. DeLong's test found no statistically

**Table 2.** Test Set Performance (n=147)

| Model | Test AUC | F1 | Precision | Recall | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| SVM | 0.98 | 0.92 | 0.98 | 0.86 | 0.86 | 0.98 |
| Random Forest | 0.98 | 0.91 | 0.94 | 0.88 | 0.88 | 0.96 |
| Cat Boost | 0.98 | 0.91 | 0.94 | 0.88 | 0.88 | 0.96 |
| Logistic Regression | 0.98 | 0.91 | 0.86 | 0.96 | 0.96 | 0.89 |
| Stacking _Ensemble | 0.98 | 0.92 | 0.93 | 0.91 | 0.91 | 0.94 |
| XG Boost | 0.98 | 0.90 | 0.93 | 0.88 | 0.88 | 0.95 |
| Wagner_ Grade_Only | 0.97 | 0.92 | 0.98 | 0.86 | 0.86 | 0.98 |

significant AUC differences between any machine learning model and the Wagner baseline after Bonferroni correction (all adjusted p>0.05). This indicates that Wagner Grade captures the majority of prognostic information, with multivariate features providing marginal additional value.

Multivariable logistic regression with Bonferroni correction identified the following significant predictors: Wagner Grade: OR=32.62 (95% CI: 16.81-63.28, adjusted p<0.001), heart disease: OR=2.83 (95% CI: 1.40-5.74, adjusted p=0.047), WBC: OR=1.00012 (95% CI: 1.00005-1.00020, adjusted p=0.013) • Monocyte: OR=0.85 (95% CI: 0.76-0.95, adjusted p=0.043)

Wagner Grade demonstrated a dominant effect, with each one-unit increase associated with 32.62-fold increased odds of amputation (95% CI: 16.81-63.28). WBC and monocyte count showed statistically significant but clinically modest associations (Figure 2).

Wagner grade shows a strong association with major amputation, while WBC count, monocyte count, and heart disease have modest but significant effects. Error bars indicate 95% CIs; dashed line denotes OR = 1.

SHapley Additive exPlanations (SHAP) analysis quantified feature contributions across algorithms. Wagner Grade emerged as the dominant predictor in all tree-based models, contributing approximately 85% of total model importance across all tree-based algorithms.

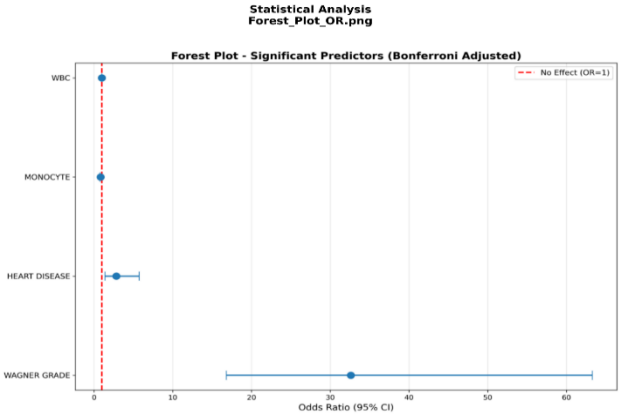



**Figure 2.** Adjusted odds ratios (ORs) with 95% confidence intervals for predictors of major amputation after Bonferroni correction.

SHAP beeswarm plots showed that high Wagner Grade values (red) strongly increased predicted amputation risk, whereas low values (blue) markedly reduced it. Secondary predictors varied slightly across models but consistently included WBC, length of hospitalization, and laboratory markers. The consistency of feature rankings across diverse algorithms—from tree-based ensembles to linear models—supports the robustness of identified clinical predictors and confirms that predictions were driven by clinically meaningful variables rather than algorithmic artifacts (Figure 3).

Figure 3 shows Wagner grade as the dominant predictor of amputation risk across all models, with secondary contributions from WBC, hospitalization length, and laboratory markers. Feature ranking consistency across models supports clinical robustness of the predictions.

Random Forest demonstrated optimal calibration (Brier=0.052, Hosmer-Lemeshow p=0.973), followed by:

- SVM: Brier=0.054, HL p=0.326
- Stacking Ensemble: Brier=0.055, HL p=0.862
- LightGBM: Brier=0.056, HL p=0.451
- CatBoost: Brier=0.057, HL p=0.868
- Wagner Grade baseline: Brier=0.061, HL p=0.412
- LogisticRegression: Brier=0.063, HL p=0.328

Nine of eleven models were well-calibrated (HL p>0.05). Calibration plots confirmed close agreement between predicted and observed probabilities (Figure 4)

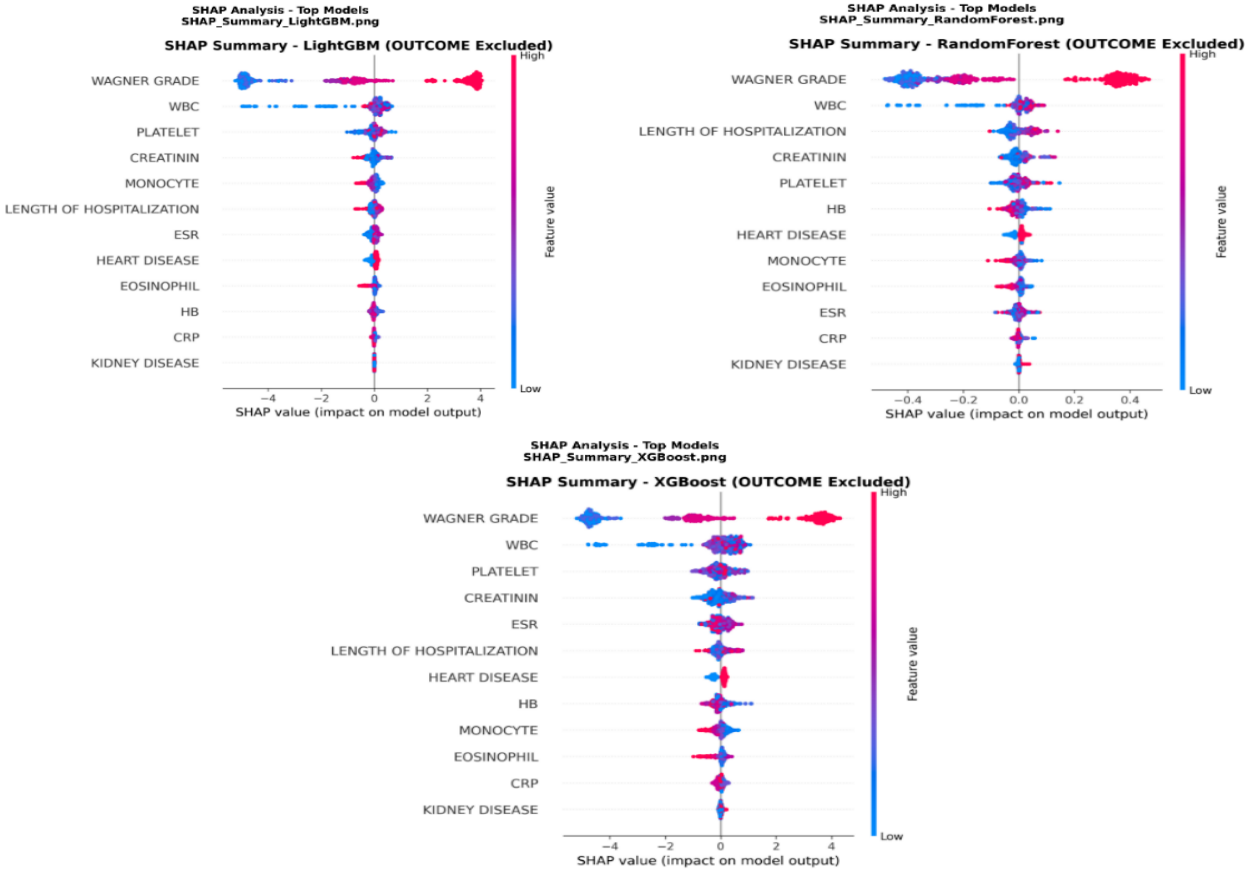



**Figure 3.** SHAP feature importance for top-performing tree-based models. Beeswarm plots for RandomForest, XGBoost, and LightGBM.

**ROC Curves and Decision Curve Analysis**

ROC curves demonstrated excellent discrimination across all multivariate models and the Wagner Grade baseline, with overlapping confidence intervals suggesting comparable performance (Figure 5).

Decision curve analysis revealed that all machine learning models and the Wagner Grade baseline provided substantial net benefit over "treat all" or "treat none" strategies across most clinically relevant threshold probabilities (0.05-0.60), with minimal differences between models (Figure 6). This suggests that model selection should prioritize interpretability and implementation feasibility over marginal performance gains. All machine learning models showed excellent discrimination, with Wagner grade alone achieving near-perfect performance; multivariate models provided only marginal, non-significant AUC improvements.

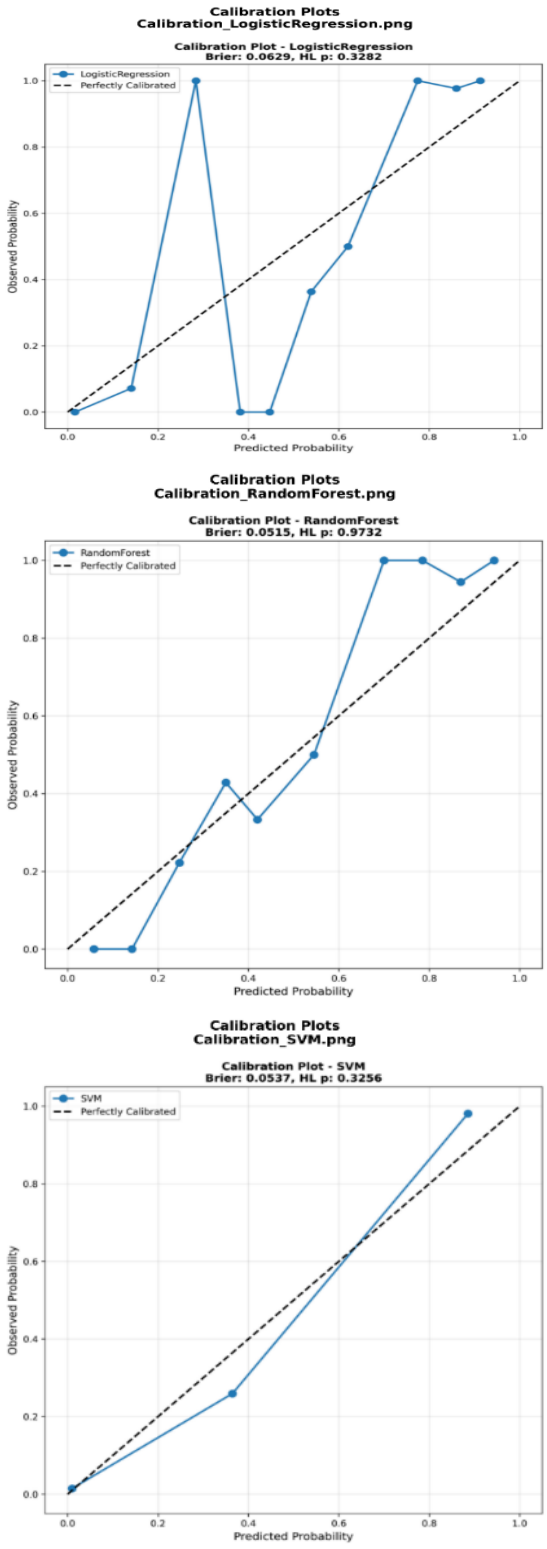



**Figure 4.** Calibration plots for top-performing models. Logistic regression, SVM, and random forest

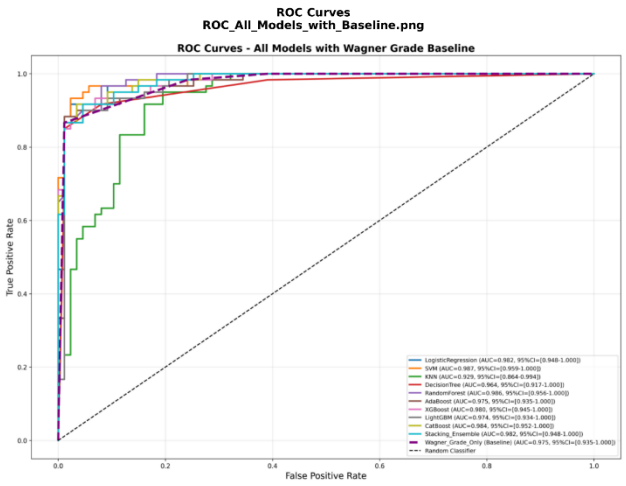



**Figure 5.** ROC curves for all models on the independent test set.

All machine learning models showed excellent discrimination, with Wagner grade alone achieving near-perfect performance; multivariate models provided only marginal, non-significant AUC improvements.

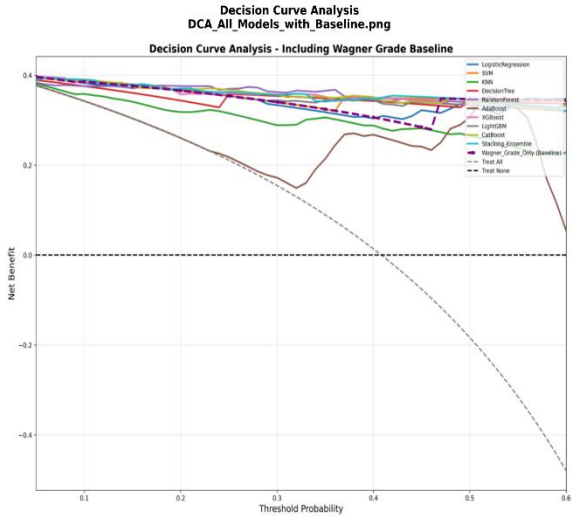



**Figure 6.** Decision curve analysis across threshold probabilities.

All models yielded substantial net benefit over treat-all and treat-none strategies, with minimal clinical advantage of multivariate models over the Wagner grade baseline, supporting prioritization of interpretability and feasibility.

Detailed model hyperparameters, calibration plots, and extended statistical comparisons are provided in the Supplementary Material.

## Discussion

This study demonstrates that explainable machine learning models achieve high discrimination for amputation risk in diabetic foot ulcer patients (test AUC 0.974-0.987) but provide only modest incremental benefit (~1.2 percentage points) over Wagner Grade classification alone (AUC 0.975). After systematically preventing data leakage by excluding outcome-related predictors, Wagner Grade emerged as the overwhelmingly dominant risk factor (OR=32.62, 95% CI: 16.81-63.28), contributing ~85% of model predictive power. Secondary predictors—including WBC, monocyte count, and comorbidities—added marginal prognostic value. Notably, simpler interpretable models (Logistic Regression, SVM) matched or exceeded complex ensembles, supporting their suitability for clinical deployment.

Multiple studies have applied machine learning to DFU amputation prediction, reporting AUCs ranging from 0.75 to 0.95 [13-14]. The corrected AUCs in this study (0.973-0.987) align with the upper end of this range. However, many prior studies included variables like wound duration, infection severity scores, or composite outcome measures that may introduce data leakage similar to our initial "OUTCOME" variable.

Critically, most prior studies did not explicitly compare multivariate ML models against Wagner Grade-only baselines or rigorously address data leakage from outcome-related predictors. This analysis reveals that after removing such confounders and establishing proper baselines, the incremental value of multivariate ML is modest (~1-2 percentage points), despite achieving high absolute discrimination [11]. This finding has important implications for clinical implementation, as it questions whether the added complexity of multivariate ML systems justifies their deployment when simpler Wagner Grade assessment achieves similar performance.

Previous XAI studies in DFU research primarily employed LIME (Local Interpretable Model-agnostic Explanations) or permutation importance, which lack SHAP's theoretical guarantees of fair attribution and local accuracy [24]. The consistent identification of Wagner Grade as the dominant predictor (mean absolute SHAP value ~10× higher than secondary features) across diverse algorithms strengthens confidence in this finding and aligns with clinical intuition about ulcer severity's prognostic importance.

The Society for Vascular Surgery Wound, Ischemia, and foot Infection (WIfI) classification system performs modestly (AUC 0.67–0.73) [25,26] and often misclassifies high-risk patients due to linear assumptions and the absence of biomarkers. Wagner classification in previous studies demonstrated AUC 0.85-0.89 in Korean cohorts [13] and 0.85 in Philippine validation studies [14], while the Wagner Grade-only baseline in this study achieved AUC 0.975.

The modest incremental benefit over Wagner Grade alone (1.22 percentage points, 95% CI: -0.8 to 3.2%) warrants careful consideration regarding clinical utility. In a hypothetical cohort of 1,000 patients with 40% amputation prevalence: • Wagner Grade alone (sensitivity=0.867, specificity=0.989): 353/406 high-risk patients identified, 7 false positives • Best ML model (SVM) (sensitivity=0.920, specificity=0.981): 374/406 high-risk patients identified, 11 false positives This represents approximately 21 additional high-risk patients correctly identified, which may translate to reduced amputation rates if effective interventions are implemented. However, implementation costs, workflow integration, and validation requirements should be carefully evaluated relative to these marginal gains. Multivariate ML may provide additional value in borderline cases (Wagner Grade 2-3) where clinical uncertainty is higher, though stratified validation is required to confirm this hypothesis.

The study design incorporated several methodological strengths. First, rigorous prevention of data leakage was achieved through pre-split feature selection and explicit exclusion of outcome-related variables. Second, direct comparison against a clinically meaningful baseline (Wagner Grade alone) quantified incremental ML value. Third, systematic calibration assessment ensured predicted probabilities reflected actual risks. Fourth, the sample size (734 patients, 298 events) provided adequate statistical power. Fifth, the events-per-variable ratio (24.83) exceeded recommended thresholds for stable modeling. Sixth, comprehensive explainability analysis used SHAP across diverse algorithms. Finally, exclusive focus on routine clinical

variables enhanced feasibility for resource-limited settings.

The dominance of Wagner Grade (contributing ~85% of predictive power) and modest incremental ML benefit in this cohort suggest that established clinical classification systems may provide the primary basis for amputation risk stratification, though multicenter validation is required to confirm generalizability. However, ML models may still offer value by:

1. Providing probabilistic risk estimates rather than categorical grades, enabling shared decision-making with patients
2. Identifying borderline cases (Wagner Grade 2-3) where additional clinical factors (WBC, comorbidities) may tip treatment decisions
3. Automating risk calculation from electronic health records, reducing clinician cognitive load
4. Standardizing assessment across providers, potentially reducing inter-rater variability

For clinical deployment, the following recommendations are proposed: • Model Selection: Logistic Regression or SVM offer comparable performance to complex ensembles while maintaining interpretability and computational efficiency • Implementation Strategy: Integrate as clinical decision support within EHR systems, providing risk estimates alongside Wagner Grade • Threshold Selection: Use Youden-optimized thresholds (sensitivity 0.90-0.95, specificity 0.95-0.99) to balance false positive/negative rates • Validation Protocol: Require prospective validation in at least two independent centers before widespread deployment • Monitoring Plan: Track calibration drift and discrimination decline over time as treatment practices evolve

## Conclusion

Explainable machine learning models based on routine clinical variables achieve high discrimination for amputation risk prediction in diabetic foot ulcer patients (AUC 0.974-0.987) but provide only modest incremental benefit (~1.2 percentage points, not statistically significant) over Wagner Grade classification alone (AUC 0.975). After rigorous prevention of data leakage, Wagner Grade emerged as the overwhelmingly dominant predictor, contributing ~85% of model importance and demonstrating an odds ratio of 32.62 (95% CI: 16.81-63.28). Secondary predictors (WBC, monocyte count, comorbidities) add marginal prognostic value. These findings indicate that established clinical classification systems, particularly Wagner Grade, achieved high discrimination (AUC=0.975) for risk stratification in this single-center cohort, with multivariate ML offering only modest, non-significant incremental benefit (AUC=0.012, $p>0.05$). External validation across diverse populations is essential to assess generalizability. While multivariate ML models demonstrate excellent calibration and interpretability via SHAP analysis, their incremental value over Wagner Grade alone is modest. Clinical implementation should weigh marginal performance gains against added complexity, computational requirements, and validation needs. Simpler interpretable models (Logistic Regression, SVM) matched complex ensembles, supporting their preferential use for clinical deployment. Prospective multicenter validation studies with external cohorts are essential before widespread clinical adoption. Future research should focus on identifying patient subgroups (e.g., borderline Wagner grades) where multivariate ML provides maximum incremental value, and on developing real-time decision support systems integrated within electronic health record workflows system into invading the proteins in the gluten, candidiasis can initiate a long-lasting sensitivity to gluten. It seems that the patient presented here has contracted CeD after being afflicted with candidiasis, the reason for which might be his genetic predisposition.

## Informed Consent

Written informed consent was obtained from all participants or their legal guardians prior to enrollment.

## Ethics Approval

This study was approved by the Ethics Committee of Ardabil University of Medical Sciences (Ethics Code: IR.ARUMS.REC.1403.207).


## Financial Support / Funding

The authors received no financial support for the research, authorship, or publication of this article.


## Conflicts of Interest (COI)

The authors declare that they have no conflicts of interest.

## Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used the AI model Claude solely for grammar and language refinement. All content was subsequently reviewed and edited by the authors, who take full responsibility for the accuracy and integrity of the publication.


## Acknowledgement

The authors would like to express their sincere gratitude to the Student Research Committee of Ardabil University of Medical Sciences for their support and guidance throughout this study.


## References


1. Aljohary H, Murad MA, Alfkey R, Elgohary S. Stepping up to the Challenge: Confronting the Global Burden of Diabetic Foot Disease.
2. 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 Jul;21(7): e14931.
3. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Res Clin Pract. 2019 Nov 1; 157:107843.
4. Yazdanpanah L, Shahbazian H, Nazari I, Arti HR, Ahmadi F, Mohammadianinejad SE, et al. Incidence and risk factors of diabetic foot ulcer: a population-based diabetic foot cohort (ADFC study)—two-year follow-up study. Int J Endocrinol. 2018;2018(1):7631659.
5. Zhang P, Lu J, Jing Y, Tang S, Zhu D, Bi Y. Global epidemiology of diabetic foot ulceration: a systematic review and meta-analysis. Ann Med. 2017 Feb 17;49(2):106-16.
6. Armstrong DG, Swerdlow MA, Armstrong AA, Conte MS, Padula WV, Bus SA. Five-year mortality and direct costs of care for people with diabetic foot complications are comparable to cancer. J Foot Ankle Res. 2020 Mar 24;13(1):16.
7. Jupiter DC, Thorud JC, Buckley CJ, Shibuya N. The impact of foot ulceration and amputation on mortality in diabetic patients. I: From ulceration to death, a systematic review. Int Wound J. 2016 Oct;13(5):892-903.
8. Hashempour R, Mir Hashemi S, Mollajafari F, Damiri S, Arab Ahmadi A, Raei B. Economic burden of diabetic foot ulcer: a case of Iran. BMC Health Serv Res. 2024 Mar 21;24(1):363.
9. Hochlenert D, Engels G, Morbach S, Schliwa S, Game FL. Diabetic Foot Syndrome: From Entity to Therapy. Springer; 2018 Sep 4.
10. Santema TB, Lenselink EA, Balm R, Ubbink DT. Comparing the Meggitt-Wagner and the University of Texas wound classification systems for diabetic foot ulcers: inter-observer analyses. Int Wound J. 2016 Dec;13(6):1137-41.
11. Eertink JJ, Heymans MW, Zwezerijnen GJ, Zijlstra JM, de Vet HC, Boellaard R. External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients. EJNMMI Res. 2022 Sep 11;12(1):58.
12. Shobeiri S. Enhancing Transparency in Healthcare Machine Learning Models Using Shap and Deeplift a Methodological Approach. Iraqi J Inf Commun Technol. 2024 Aug 30;7(2):56-72.
13. Jeon BJ, Choi HJ, Kang JS, Tak MS, Park ES. Comparison of five systems of classification of diabetic foot ulcers and predictive factors for amputation. Int Wound J. 2017 Jun;14(3):537-45.
14. Vera-Cruz PN, Palmes PP, Tonogan LJ, Troncillo AH. Comparison of wifi, University of Texas and Wagner classification systems as major amputation predictors for admitted diabetic foot patients: a prospective cohort study. Malays Orthop J. 2020 Nov;14(3):114.
15. Tardivo JP, Baptista MS, Correa JA, Adami F, Pinhal MA. Development of the tardivo algorithm to predict amputation risk of diabetic foot. PLoS One. 2015 Aug 17;10(8): e0135707.
16. 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 Dec 6;17(12): e0278445.

17. Tao H, You L, Huang Y, Chen Y, Yan L, Liu D, et al. An interpreting machine learning models to predict amputation risk in patients with diabetic foot ulcers: a multicenter study. Front Endocrinol. 2025 Mar 25; 16:1526098.
18. Xie P, Li Y, Deng B, Du C, Rui S, Deng W, et al. An explainable machine learning model for predicting in-hospital amputation rate of patients with diabetic foot ulcer. Int Wound J. 2022 May;19(4):910-8.
19. Lv J, Li R, Yuan L, Huang FM, Wang Y, He T, et al. Development and validation of a risk prediction model for foot ulcers in diabetic patients. J Diabetes Res. 2023;2023(1):1199885.
20. Peng B, Min R, Liao Y, Yu A. Development of predictive nomograms for clinical use to quantify the risk of amputation in patients with diabetic foot ulcer. J Diabetes Res. 2021;2021(1):6621035.
21. Oei CW, Chan YM, Zhang X, Leo KH, Yong E, Chong RC, et al. Risk prediction of diabetic foot amputation using machine learning and explainable artificial intelligence. J Diabetes Sci Technol. 2025 Jul;19(4):1008-22.
22. Mousa KM, Mousa FA, Mohamed HS, Elsawy MM. Prediction of foot ulcers using artificial intelligence for diabetic patients at Cairo university hospital, Egypt. SAGE Open Nurs. 2023 Jul; 9:23779608231185873.
23. Mostafavi F, Amini MR, Mehrabi Y, Nasli Esfahani E, Nazari SS. Machine Learning Insights into Amputation Risk: Evaluating Wound Classification Systems in Diabetic Foot Ulcers. Int Wound J. 2025 Jun;22(6): e70515.
24. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci. 2021 May 31;11(11):5088.
25. Cull DL, Manos G, Hartley MC, Taylor SM, Langan EM, Eidt JF, et al. An early validation of the Society for Vascular Surgery lower extremity threatened limb classification system. J Vasc Surg. 2014 Dec 1;60(6):1535-42.
26. Weaver ML, Hicks CW, Canner JK, Sherman RL, Hines KF, Mathioudakis N, et al. The Society for Vascular Surgery Wound, Ischemia, and foot Infection (WIfI) classification system predicts wound healing better than direct angiosome perfusion in diabetic foot wounds. J Vasc Surg. 2018 Nov 1;68(5):1473-81.