Original article

An Advanced Machine Learning Framework to Identify Associated Factors and Predict the Risk of Oral Candidiasis in Cancer Patients

Bach Duy Thai[1], Truong Tuan Khanh[1], Bui Anh Thu[1], Trieu Thi Oanh[1], Bui Tuan Dat[1], Nguyen Thi Hong Chuyen[2], Ton Nu Phuong Anh[3*]

[1]Medical Student, 2020–2026, University of Medicine and Pharmacy, Hue University, Hue, Vietnam

[2]Department of Oncology, University of Medicine and Pharmacy, Hue University, Hue, Vietnam

[3]Department of Parasitology, University of Medicine and Pharmacy, Hue University, Hue, Vietnam

Abstract

Background: Oral candidiasis is a common opportunistic infection in cancer patients, particularly those undergoing chemotherapy. Multiple clinical and hematological factors contribute to infection risk, but their complex interactions remain poorly understood using conventional statistical methods. **Objectives:** To identify associated factors, and develop machine learning models to predict infection risk of oral candidiasis among cancer patients receiving and not yet receiving chemotherapy. **Materials and Methods:** This cross-sectional study enrolled 69 cancer patients at Hue University of Medicine and Pharmacy Hospital between October 2024 and May 2025. Patients underwent clinical examinations, laboratory testing, direct oral swab microscopy and cultivation for candidiasis diagnosis. Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) was used to select relevant features. eXtreme Gradient Boosting (XGBoost) models were developed for each patient group (chemotherapy and non-chemotherapy) and interpreted using SHapley Additive exPlanations (SHAP) value method. **Results:** Oral candidiasis was detected in 36.8% of chemotherapy patients and 35.4% of non-chemotherapy patients. Key associated factors included dry mouth, taste change, white patches on mucosa, low lymphocyte or red blood cell counts, poor oral hygiene, and antibiotic use. XGBoost models achieved high performance in both groups (AUC-ROC: 0.9093 for chemotherapy; 0.8758 for non-chemotherapy). SHAP analysis revealed feature-specific contributions aligned with clinical relevance, confirming the model's interpretability and consistency. **Conclusion:** Oral candidiasis is highly prevalent among cancer patients, with distinct risk profiles between those with and without chemotherapy. Machine learning methods such as sPLS-DA and XGBoost effectively identified and interpreted predictive factors, offering valuable tools for clinical risk stratification and early prevention in oncology care.

Keywords: *Oral Candidiasis; Cancer; Chemotherapy; Machine Learning; sPLS-DA; XGBoost.*

1. INTRODUCTION

Candida species, particularly *Candida albicans*, are commonly found in the oral cavity as part of the normal microbiota, but can become opportunistic pathogens under conditions of immunosuppression, leading to oral candidiasis [1]. This pathogenic transition is often triggered by local or systemic disturbances, especially in vulnerable populations such as cancer patients [1]. Cancer therapies, especially chemotherapy and radiotherapy, can compromise immune function, damage the oral mucosa, and disrupt microbial homeostasis, thereby promoting *Candida* overgrowth [1, 2, 4, 5]. Oral candidiasis is a frequent complication in this setting, with prevalence rates reaching 39.1% during cancer treatment and up to 53.5% in patients receiving head and neck radiotherapy [2, 4]. Clinical features such as oral burning, dysgeusia, dysphagia, and mucosal patches negatively affect quality of life, nutrition, and may even lead to systemic candidemia or delay cancer therapy [1, 2, 4]. Moreover, differentiating candidiasis from radiation- or chemotherapy-induced mucositis remains a diagnostic challenge [4].

Traditional statistical methods often struggle to elucidate the complex and potentially nonlinear relationships among multiple risk factors, potentially overlooking latent structures or rare yet important contributors [6, 7]. In contrast, machine learning offers advanced tools capable of processing complex datasets, automatically uncovering hidden patterns, and modeling multidimensional relationships without strict a priori assumptions [6, 7]. Such techniques have been successfully applied in medicine to predict disease risk and identify novel

**Corresponding Author: Ton Nu Phuong Anh; tnpanh@huemed-univ.edu.vn*

DOI: 10.34071/jmp.2025.6.889

Received: 30/06/2025; Accepted: 25/10/2025; Published: 30/12/2025

predictors, including for *Candida* infections [6, 7].

Although several risk factors for oral candidiasis in cancer patients have been proposed, the intricate interactions, particularly those involving multimodal therapies or detailed behavioral habits remain insufficiently clarified through conventional approaches. Many prior studies are limited in scope or sample size. Our study, titled "An Advanced Machine Learning Framework to Identify Associated Factors and Predict the Risk of Oral Candidiasis in Cancer Patients", aimed to address these gaps through two main objectives:

1. To assess factors statistically associated with oral candidiasis in patients with and without chemotherapy.

2. To develop and evaluate an XGBoost model for predicting oral candidiasis risk among cancer patients.

2. MATERIALS AND METHODS

2.1. Subjects

Eligible participants were adult patients with a confirmed cancer diagnosis established by oncology specialists, based on histopathological findings and relevant imaging studies.

Patients were excluded if they refused to provide relevant information; had a history of oral mucosal infections unrelated to Candida (e.g., Herpes simplex or other ulcerative conditions); were diagnosed with immunodeficiency disorders (including HIV/AIDS, systemic lupus erythematosus, or congenital immunodeficiencies); or were unable to complete clinical assessments and laboratory procedures due to psychiatric, physical, or other limiting conditions.

2.2. Research Methods

2.2.1. Study Design and Sample Size

This cross-sectional study was conducted from October 2024 to May 2025 at the Department of Oncology and the Department of Parasitology, Hue University of Medicine and Pharmacy. A total of 69 patients were recruited using a convenience sampling method.

2.2.2. Data Collection

Patient data were collected through direct interviews and medical record reviews. Collected variables included demographic characteristics (gender, age, BMI), medical history (comorbidities, smoking status, ≥20 pack-years, recent antifungal use), and current clinical condition (presence of concurrent infections). Cancer-related data encompassed cancer type, chemotherapy status, number of cycles and class of chemotherapy agents, other concurrent therapies, and relevant risk factors such as prolonged hospitalization, central venous catheterization, total parenteral nutrition, and dialysis. Oral hygiene practices (brushing frequency, post-meal oral care, denture use) and laboratory values (complete blood count with detailed leukocyte differentials, red blood cells, and platelets) were also recorded. Each patient underwent a focused oral examination to identify symptoms (burning mouth, taste change, anorexia, dry mouth) and clinical signs (white patches, angular cheilitis, ulcers, erythema, smooth or nodular tongue, depapillated or black hairy tongue, halitosis).

2.2.3. Sample Collection

Oral swabs were collected in the morning using sterile cotton swabs after oral hygiene. Samples were immediately transferred to the Department of Parasitology, Hue University of Medicine and Pharmacy Hospital, for same-day analysis.

2.2.4. Laboratory Testing

Oral swab specimens were treated with potassium hydroxide (KOH) and examined under light microscopy at 40× magnification to detect fungal elements. Subsequently, all samples were cultured on Sabouraud dextrose agar supplemented with chloramphenicol for fungal isolation. *Candida albicans* and *Candida non-albicans* species were identified using chromogenic agar. A diagnosis of oral candidiasis was established based on the isolation of *Candida* species from culture.

2.2.5. Data Analysis

2.2.5.1. Data Preprocessing

All variables were entered into SPSS version 27. Categorical variables were binarized using one-hot encoding. No significant outliers were identified upon inspection.

2.2.5.2. Data Stratification

The dataset was divided into two main groups:

- Chemotherapy group: patients who had completed at least one chemotherapy cycle.
- Non-chemotherapy group: patients who have not yet received chemotherapy.

2.2.5.3. Data Exploration

Distribution was assessed using skewness, kurtosis, and Shapiro–Wilk test for normality. Categorical variables were summarized using frequencies and percentages. Continuous variables were reported as mean ± SD for normally distributed data, and median (interquartile range) for non-normally distributed data.

Group comparisons between chemotherapy and non-chemotherapy groups were conducted using

Chi-square test or Fisher's exact test for categorical variables, independent samples t-test for normally distributed continuous variables and Mann–Whitney U test for non-normally distributed continuous variables. A p-value < 0.05 was considered statistically significant.

Further analyses were conducted separately for the chemotherapy and non-chemotherapy groups:

2.2.5.4. Feature selection

Discriminant analysis was conducted using sparse Partial Least Squares Discriminant Analysis (sPLS-DA) implemented via the MixOmics package through the rpy2 interface in Python 3.13. The analysis incorporated all collected variables, including anthropometric characteristics, medical history, clinical signs and symptoms, laboratory results, cancer type, and treatment modalities. Prior to modeling, the dataset was preprocessed with Square Root transformation followed by Auto Scaling (mean-centering and variance-scaling). The performance of the sPLS-DA model was evaluated using the perf function with 5-fold cross-validation repeated 10 times, in order to assess classification accuracy and model stability.

The variables selected from the sPLS-DA analysis were re-evaluated using the same statistical tests described in the data exploration step.

A correlation matrix was subsequently constructed to assess multicollinearity among these variables. Pairs of variables exhibiting a strong correlation ($|r| \geq 0.8$) were considered for exclusion.

2.2.5.5. Model Training

The selected variables identified in the previous steps were used to train the XGBoost (eXtreme Gradient Boosting). Those variables were imported into Python 3.13 using the pandas library. To address class imbalance between *Candida*-positive and *Candida*-negative groups, the scale_pos_weight parameter in the XGBoost model was calculated based on the actual class distribution of the target variable.

2.2.5.6. Model Optimization

An initial XGBoost model was trained using a nested stratified K-Fold cross-validation approach (k=5 for the chemotherapy group and k = 4 for the non-chemotherapy group), repeated 20 times.

Hyperparameter tuning was performed using GridSearchCV within each fold. The best-performing parameters were then used to retrain the model using the same nested stratified K-Fold scheme (k = 5 or k = 4, repeated 20 times).

After hyperparameter optimization, the classification threshold was adjusted based on Youden's J statistic derived from the ROC curve of each fold. The final model was retrained using both the optimized hyperparameters and thresholds under the same repeated nested stratified K-Fold scheme.

2.2.5.7. Model Evaluation

The performance of the initial model, the hyperparameter-optimized model, and the threshold-optimized model was compared. For each model, performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC were calculated across all folds and summarized as mean values, standard deviations, and 95% confidence intervals, presented in a comparative summary table.

2.2.5.8. Model Interpretation

SHAP (SHapley Additive exPlanations) values were computed for the threshold-optimized XGBoost model to interpret the contribution of individual variables to model predictions. Visualization of SHAP values was performed using the matplotlib library.

3. RESULTS

3.1. Baseline characteristics of Study Participants

Gastrointestinal cancer was the most common type, accounting for 30.4% in the non-chemotherapy group and 42% in the chemotherapy group. Surgery was the most common adjunctive treatment (23 patients). Oral hygiene practices were similar between groups. Median toothbrushing frequency was 1 time/day (IQR: 1 - 2; p = 0.933). Rinsing or brushing after meals was reported by 31.9% of non-chemotherapy and 27.5% of chemotherapy patients (p = 0.078). No patients used dentures. Among chemotherapy patients (N = 38), 31.6% in the oral candidiasis group and 50.0% in the non-oral candidiasis group received two or more agents (p = 1.000), with no significant differences in chemotherapeutic classes used (p ≥ 0.433) or median number of cycles (2 cycles, p = 0.709).

Oral candidiasis was diagnosed in 11 non-chemotherapy patients (15.9%) and 14 chemotherapy patients (20.3%), with Candida positivity rates of 35.4% and 36.8%, respectively.

Table 1. Clinical and Laboratory Characteristics of the Study Population.

Baseline characteristics		Non chemotherapy (N = 31)		Chemotherapy (N =3 8)		p
		N	%	N	%	
Symptoms and signs						
Symptoms	Oral burning pain	6	8.7	2	2.9	0.127*
	Taste change, anorexia	8	11.6	18	26.1	0.066
	Dry mouth	13	18.8	19	27.5	0.504
	Asymptomatic	11	15.9	15	21.7	0.734
Signs	White patches on mucosa	11	15.9	14	20.3	0.907
	Redness at mouth corners	5	7.2	8	11.6	0.603
	Red inflamed oral mucosa	2	2.9	3	4.3	1*
	Glossy tongue or small papillae	1	3.2	4	5.8	0.370*
	Erythematous depapillated tongue	1	1.4	0	0	0.449*
	Halitosis	12	17.4	13	18.8	0.699
	No signs	12	17.4	15	21.7	0.948
Blood cell count in complete blood count (G/L)						
White blood cell count	Median (IQR)	7.99 (5.33 - 10.66)		8.18 (5.84 - 10.52)		0.554
Neutrophil count	**Mean ± SD**	**9.90 ± 1.79**		**8.72 ± 2.52**		**0.032**
Lymphocyte count	**Median (IQR)**	**2.58 (1.58 - 3.58)**		**3.77 (2.27 - 5.27)**		**0.031**
Monocyte count	Mean ± SD	1.27 ± 0.41		1.38 ± 0.53		0.311
Eosinophil count	Median (IQR)	0.20 (0.07 - 0.33)		0.20 (0.00 - 0.41)		0.933
Basophil count	Median (IQR)	0.06 (0.01 - 0.10)		0.08 (0.03 - 0.13)		0.813
Red blood cell count	Median (IQR)	4.22 (3.69 - 4.75)		3.78 (3.28 - 4.29)		0.156
Platelet count	Median (IQR)	266 (206.5 - 325.5)		285 (201.0 - 369.0)		0.405
Direct microscopic and culture results						
Oral candidiasis		11	15.9	14	20.3	0.907

Note: % within total, (*) Fisher's exact test.

3.2. Investigation of Factors Associated with Oral Candidiasis in Chemotherapy Patients

Oral candidiasis and non oral candidiasis (Figure 1A), with Component 1 and Component 2 explaining 17.7% and 7.6% of the variance, respectively. Minimal overlap in the 95% confidence intervals between groups indicated good discriminative capacity.

Key variables contributing to group separation were identified based on loading values (Figure 1B). Clinical features such as taste alteration, anorexia, dry mouth, concurrent bacterial infection, white patches on the mucosa, and elevated neutrophil count were associated with infection. In contrast, higher lymphocyte and basophil counts and absence of symptoms or signs were linked to non-infected patients.

All selected variables showed statistically significant differences between groups ($p < 0.01$, Figure 1C), reinforcing their discriminatory potential.

A correlation heatmap (Figure 1D) revealed strong co-occurrence of symptoms, particularly between taste change, anorexia, dry mouth, and white patches ($r = 0.63 - 0.70$). The variable "asymptomatic" was negatively correlated with these features ($r = -0.62$ to -0.77). Hematologically, lymphocyte and basophil counts were positively correlated ($r = 0.66$) and both negatively correlated with neutrophils ($r = -0.94$ and -0.74, respectively). Concurrent bacterial infection correlated positively with neutrophils ($r = 0.65$) and negatively with lymphocytes ($r = -0.65$).

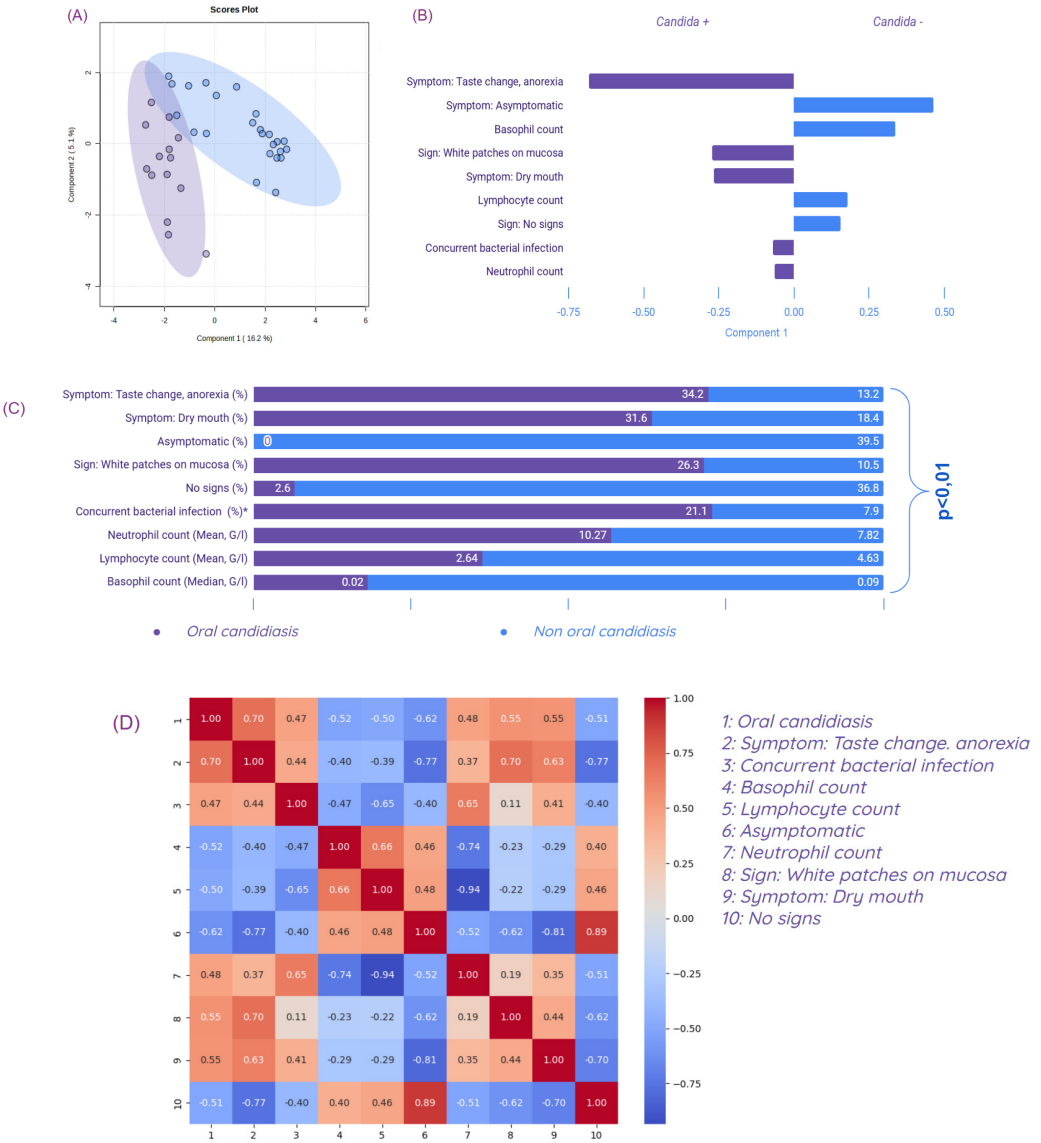

Figure 1. Feature selection in chemotherapy group. (A) sPLS-DA scatter plot showing separation of Oral candidiasis and non oral candidiasis patients. **(B)** Variable importance based on Component 1 loadings. **(C)** Distribution and statistical comparison of selected features. **(D)** Correlation matrix of key features. *Note: % within total, (*) Fisher's exact test*

Figures 2A and 2B illustrate the contribution of various features to the XGBoost model's prediction of oral candidiasis risk. The presence of white patches on the mucosa was the most influential feature, with the highest mean SHAP value (~0.215), and was strongly associated with an increased probability of *Candida*-positive classification (Figure 2B). Dry mouth was the second most influential clinical symptom (mean SHAP value ~0.152), significantly increasing the predicted probability of infection. Among hematological indices, lymphocyte count was notably impactful (mean SHAP value ~0.126), with lower values (blue) increasing the predicted risk (Figure 2B). Other features like basophil count, taste change, anorexia, concurrent bacterial infection, and neutrophil count contributed to a lesser degree. Conversely, absence of symptoms and signs had minimal or zero SHAP values, indicating their minimal contribution to positive predictions and their characteristic presence in *Candida*-negative cases.

Figures 2C and 2D demonstrate how individual features contribute to XGBoost predictions, revealing context-dependent influences. For Sample 10 (Figure

2C), the absence of white patches on mucosa (SHAP = –0.22) was the strongest negative contributor, correctly driving a negative prediction despite minor positive influences from dry mouth and elevated lymphocyte count. Conversely, in Sample 37 (Figure 2D), the absence of white patches also had a strong negative SHAP (–0.22). However, this was overpowered by strong positive contributions from dry mouth, lymphocyte count, and notably, a basophil count of 0 (which contributed positively here, unlike Sample 10), leading to a correct positive prediction. This comparison highlights how a feature's impact can reverse or be outweighed by other factors depending on the overall clinical profile.

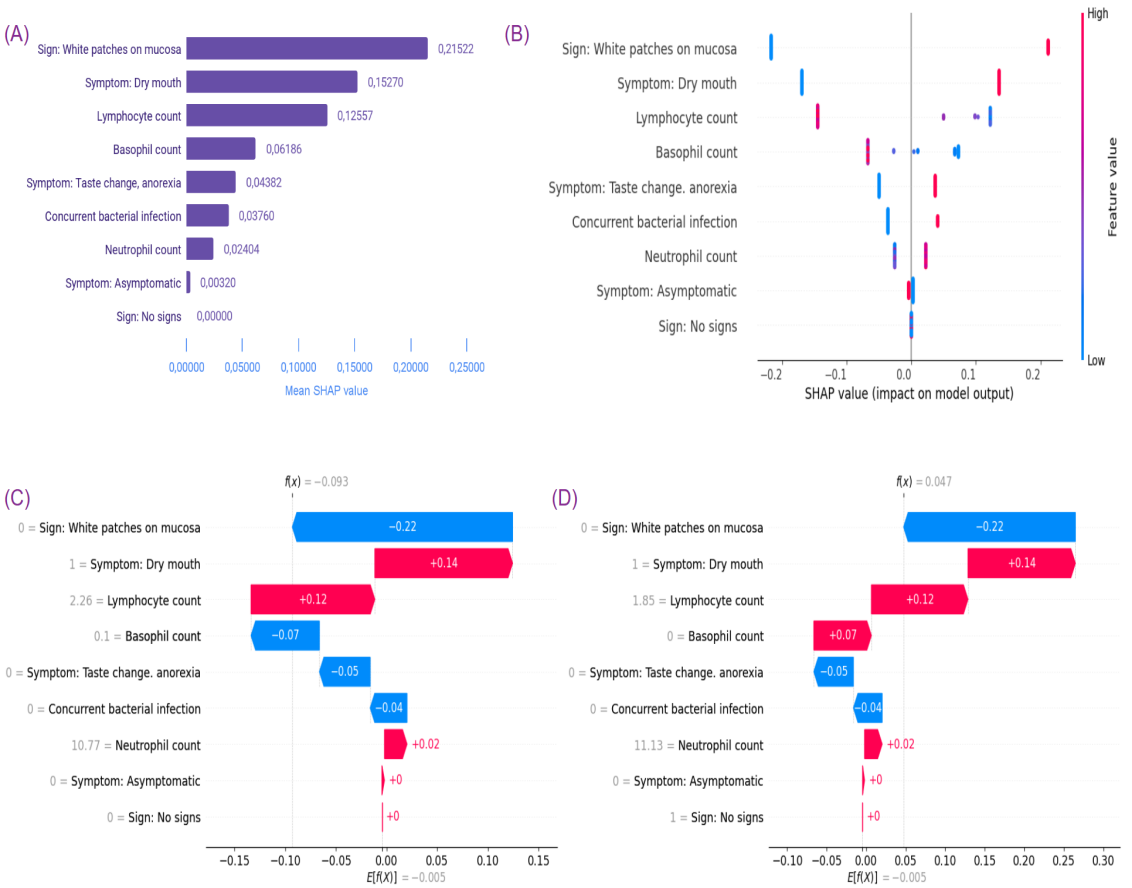

Figure 2. Model Interpretation Using SHAP. (A) SHAP summary plot showing the mean absolute SHAP values of each feature. **(B)** SHAP beeswarm plot illustrating the individual impact of each feature on the model's prediction. **(C)** SHAP waterfall plot for Sample 10 (true negative case). **(D)** SHAP waterfall plot for Sample 37 (true positive case).

3.3. Investigation of Factors Associated with Oral Candidiasis in Non-Chemotherapy Patients

sPLS-DA analysis demonstrated effective separation between Candida-positive and Candida-negative patients (Figure 3A), with Components 1 and 2 accounting for 17.2% and 8.4% of the variance, respectively.

Top contributing variables (Figure 3B) included clinical symptoms—taste change, anorexia, dry mouth, white patches, halitosis—and laboratory findings such as elevated neutrophil count and reduced lymphocyte/basophil counts. Absence of symptoms or signs was more common in the non-infected group.

Statistical analysis confirmed significant differences ($p < 0.05$ or $p < 0.01$) for all sPLS-DA-selected variables except total white blood cell count (Figure 3C). Notably, antibiotic use and poor oral hygiene were associated with infection, while post-meal oral care was more frequent in non-infected patients.

Correlation matrix (Figure 3D) revealed strong associations among key symptoms. Dry mouth was highly correlated with halitosis (r = 0.80), and white

patches correlated positively with halitosis (r = 0.66), taste changes (r = 0.64), and dry mouth (r = 0.60). In contrast, "no signs" was negatively correlated with these symptoms (r = –0.59 to –0.68).

Antibiotic use correlated positively with neutrophil count (r = 0.42) and white patches (r = 0.15), while rinsing/brushing after meals inversely correlated with dry mouth (r = –0.46). Red blood cell count showed negative correlations with dry mouth (r = –0.45) and halitosis (r = –0.52).

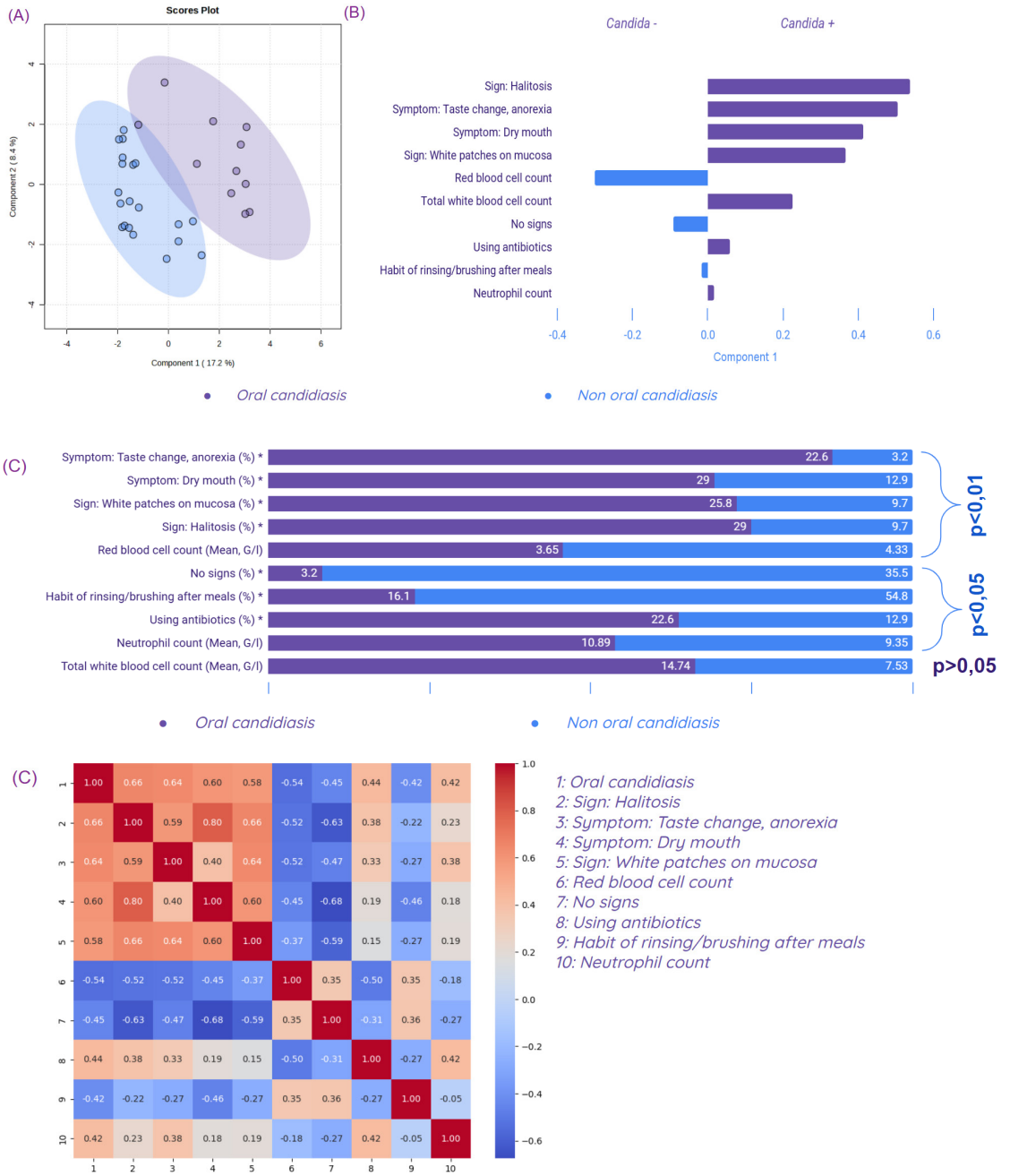

Figure 3. Feature selection in non-chemotherapy group. (A) sPLS-DA scatter plot showing separation of Oral candidiasis and non oral candidiasis patients. **(B)** Variable importance based on Component 1 loadings. **(C)** Distribution and statistical comparison of selected features. **(D)** Correlation matrix of key features. *Note: Note: % within total, (*) Fisher's exact test*

After excluding total white blood cell count, the XGBoost model was run with 9 selected variables. Figures 4A and 4B reveal halitosis as the strongest predictor of oral candidiasis in the XGBoost model, with the highest mean absolute SHAP value (-0.5862) strongly increasing predicted infection probability. Lower red blood cell counts and white patches on mucosa also showed moderate positive influence. Other features like oral hygiene, antibiotic use, dry mouth, taste change, anorexia, and absence of signs had minimal impact.

Figures 4C and 4D present SHAP waterfall plots for two cases, illustrating feature contributions to oral candidiasis predictions. Sample 9 (Figure 4C), a true-negative case, had a model output of f(x) = –0.273. Halitosis (SHAP = –0.41) was the strongest negative contributor, along with red blood cell count (SHAP = –0.16), correctly driving a negative prediction despite positive influences from white patches and neutrophil count. Sample 28 (Figure 4D), a false-negative case, showed a model output of f(x) = –0.683, incorrectly classifying an infected patient. Similar to Sample 9, halitosis (SHAP = –0.41), red blood cell count, and white patches had strong negative influences. Despite a high neutrophil count (SHAP = +0.21), it was insufficient to reverse the negative classification. This comparison highlights halitosis as a consistent negative predictor, but in Sample 28, the model underestimated positive-driving variables, leading to the false-negative outcome.

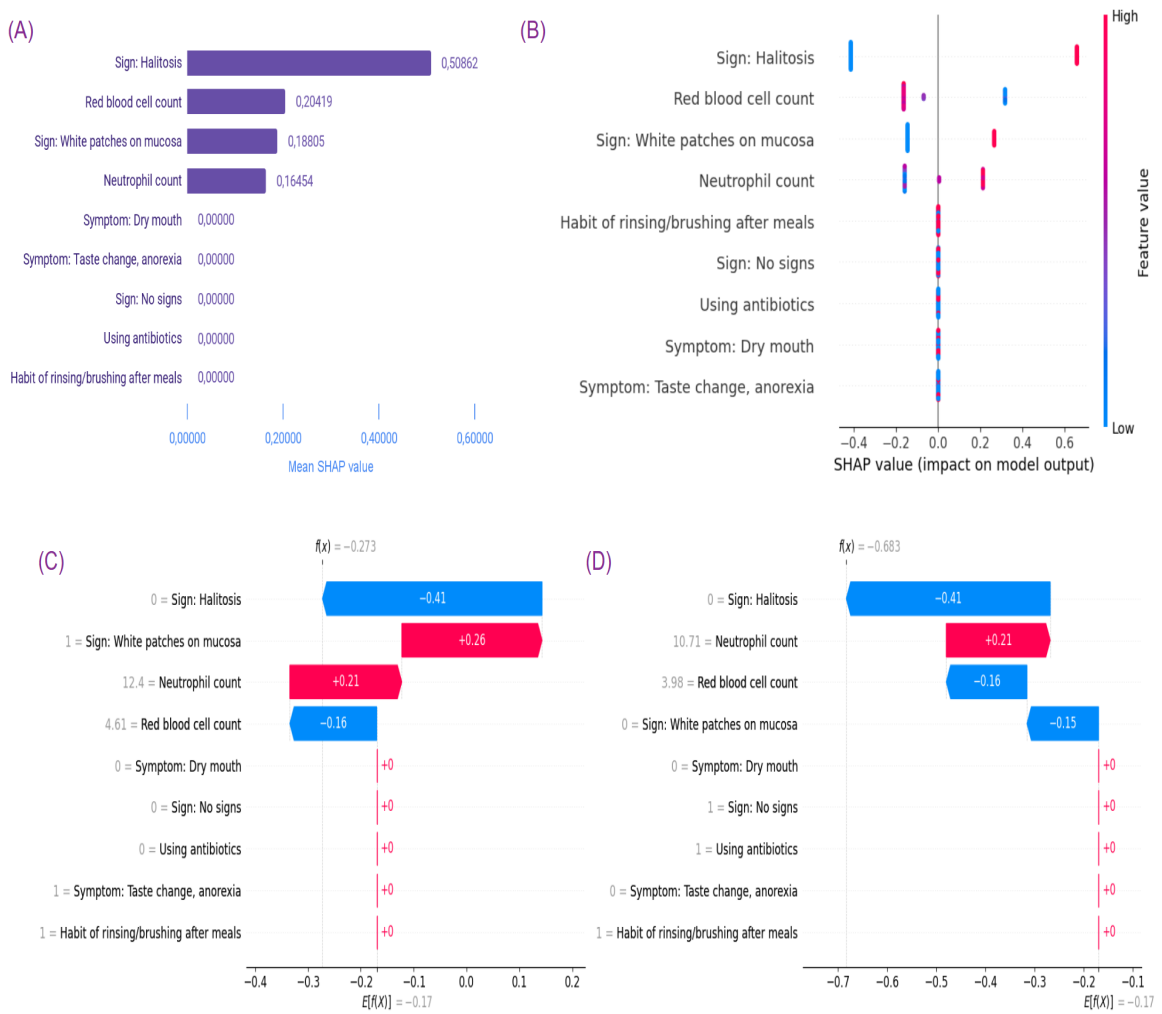

Figure 4. Model Interpretation Using SHAP. **(A)** SHAP summary plot showing the mean absolute SHAP values of each feature. **(B)** SHAP beeswarm plot illustrating the individual impact of each feature on the model's prediction. **(C)** SHAP waterfall plot for Sample 9 (true negative case). **(D)** SHAP waterfall plot for Sample 28 (false negative case).

3.4. Comparative Performance of XGBoost Models in Chemotherapy and Non- Chemotherapy Groups

In the chemotherapy group, accuracy remained consistently high across all models, with slight improvements in precision after hyperparameter tuning and a notable increase in recall following threshold optimization. The F1-score rose modestly, and AUC-ROC remained excellent.

In contrast, the non-chemotherapy group showed a slightly lower baseline accuracy, which increased after threshold optimization. Precision also improved marginally but with wider variability. Unlike the chemotherapy group, recall decreased with threshold adjustment, indicating a trade-off with improved precision. The F1-score showed a similar trend. However, AUC-ROC improved steadily, indicating enhanced discrimination overall.

When comparing the two groups, the chemotherapy group consistently achieved higher accuracy, recall, and AUC-ROC values than the non-chemotherapy group, suggesting better model performance when treatment-specific variables were included. The trade-off between precision and recall was more pronounced in the non-chemotherapy group, highlighting differences in variable contributions and prediction stability between patient cohorts.

Table 2. Comparative performance of the initial, hyperparameter-optimized, and threshold-optimized XGBoost models in chemotherapy and non-chemotherapy groups

Metrics	Chemotherapy group			Non-chemotherapy group		
	Initial model	Hyperparameter optimized model	Threshold optimized model	Initial model	Hyperparameter optimized model	Threshold optimized model
Accuracy	0.7950 ± 0.1293	0.7987 ± 0.1310	0.7975 ± 0.1344	0.7460 ± 0.1299	0.7768 ± 0.1497	0.7795 ± 0.1429
Precision	0.7190 ± 0.1892	0.7275 ± 0.1861	0.7021 ± 0.1722	0.6654 ± 0.2367	0.6994 ± 0.2799	0.7323 ± 0.3300
Recall	0.8233 ± 0.2079	0.8467 ± 0.2182	0.9333 ± 0.1491	0.6625 ± 0.2541	0.6813 ± 0.2791	0.5813 ± 0.2946
F1-score	0.7500 ± 0.1621	0.7572 ± 0.1620	0.7840 ± 0.1335	0.6361 ± 0.1984	0.6669 ± 0.2462	0.6203 ± 0.2764
AUC-ROC	0.8960 ± 0.0988	0.9093 ± 0.0924	0.9093 ± 0.0924	0.8246 ± 0.1273	0.8758 ± 0.1183	0.8758 ± 0.1183

4. DISCUSSION

Our study of 69 cancer patients found a 36.2% prevalence of oral candidiasis, which is lower than previous reports [8-10]. We observed similar prevalence rates in both non-chemotherapy (35.5%) and chemotherapy-treated (36.8%) groups, differing from some earlier findings [8]. These discrepancies might stem from our sample size, diagnostic methods, limited cancer type diversity, and varying case definitions. Despite this, our findings underscore the increased risk of fungal infection in cancer patients compared to those with internal medical patients [11]. The significant prevalence in non-chemotherapy patients also suggests that malignancy itself, or pre-existing factors, may contribute to fungal invasion.

In chemotherapy patients, local clinical features and hematologic parameters were central predictors of oral candidiasis. SHAP analysis identified white patches on the mucosa, dry mouth, and taste changes/anorexia as most influential, aligning with known clinical presentations and the impact of chemotherapy on salivary components and nutrition [1, 5, 10, 11]. Concurrent bacterial infections were also relevant, potentially due to immunosuppression or microbiome disruption [12]. Lymphopenia was a strong hematologic predictor, consistent with the role of CD4+ T cells in antifungal immunity [13]. While sPLS-DA indicated higher neutrophil counts in infected patients, SHAP analysis assigned lower importance, possibly reflecting reactive neutrophilia or functional impairment post-chemotherapy [14]. The role of basophils warrants further investigation [15].

For non chemotherapy patients, predictive variables leaned towards local and behavioral factors. Halitosis had the strongest SHAP influence, possibly linked to fungal/bacterial overgrowth [16, 17]. White patches and xerostomia also retained predictive value. Low red blood cell count showed a notable association, as anemia can compromise mucosal integrity or alter iron availability, promoting fungal growth [18, 19]. Increased neutrophil count was observed in infected patients, potentially indicating local inflammation. Behavioral factors like antibiotic use (disrupting microbiota) and poor oral hygiene were also key, reinforcing the importance of preventive care [12].

A strength of this study lies in the application of integrated machine learning approaches: sPLS-DA for variable selection and XGBoost for predictive modeling, with SHAP providing model interpretability. Both models demonstrated robust performance with AUC-ROC values > 0.87 (0.9093 in chemotherapy and 0.8758 in non-chemotherapy groups), comparable to prior studies using XGBoost in medical prediction tasks [20 - 23]. The use of SHAP allowed deeper insight into the contribution and direction of influence of each variable, improving transparency and potential clinical utility [24]. Importantly, SHAP confirmed that the most influential predictors closely matched those identified via sPLS-DA, enhancing model credibility.

Differences in key predictors between the two patient groups highlight distinct pathophysiological mechanisms: immune suppression and systemic alterations in the chemotherapy group versus local factors and hygiene-related variables in the non-chemotherapy group. These findings support the rationale for building separate models tailored to specific patient populations, as SHAP analysis reveals that the predictive importance of variables can shift across contexts.

Despite the modest sample size, we implemented a carefully designed and rigorous modeling pipeline to minimize overfitting-an inherent risk in high-dimensional, low-sample-size datasets. This included proper data preprocessing, nested stratified cross-validation with multiple repetitions, hyperparameter tuning, threshold optimization, and interpretability analysis. Our results demonstrate that even complex machine learning models can yield robust and clinically meaningful outcomes when applied with appropriate methodological safeguards. This serves as an important methodological contribution and a practical example for future studies dealing with small datasets.

The cross-sectional design precludes the establishment of causal relationships. The relatively small sample size (N = 69), particularly after subgroup stratification, reduces statistical power and limits the generalizability of the findings. This may also compromise the robustness of the machine learning models due to the high dimensionality of the data. The study was conducted at a single center, which may not reflect the broader demographic or clinical variability seen in other institutions or geographic regions. Additionally, some potentially important variables were either incompletely collected or not assessed, including detailed cancer staging, specific chemotherapy regimens, nutritional status indicators, salivary pH, and denture use.

While machine learning has rapidly expanded across numerous disciplines in recent years, its application in clinical research in Vietnam remains limited. To the best of our knowledge, this is one of the first studies in Vietnam to integrate advanced machine learning algorithms, specifically the combination of sPLS-DA, XGBoost, and SHAP for the analysis and prediction of clinical outcomes. Globally, this also represents one of the pioneering applications of advanced machine learning in the investigation of oral candidiasis, a condition that has received relatively little attention in predictive modeling research. At present, there is no strong evidence to support universal screening for oral candidiasis across all oncology populations, and such an approach would not be feasible. Our findings suggest that targeted, risk-based strategies may be more clinically relevant. Future multicenter, prospective studies with larger cohorts, additional variables, and advanced diagnostics such as PCR are needed to strengthen the validity of machine learning models while enabling the creation of external validation cohorts to enhance their reliability and generalizability. Ultimately, these advances may facilitate the development of practical clinical tools such as web-based calculators, mobile applications, or simplified risk scores that support oncologists in stratifying patients by infection risk. Such tools would allow clinicians to focus diagnostic testing on high-risk individuals, enabling earlier detection and timely preventive or therapeutic interventions.

5. CONCLUSION

In this study, the prevalence of oral *Candida* infection was high in both chemotherapy (36.8%) and non-chemotherapy (35.4%) cancer patient groups. Using sPLS-DA, we identified key discriminatory variables, including local clinical signs (e.g., dry mouth, white patches), hematological indices (e.g., lymphocyte and red blood cell counts), and background factors (e.g., antibiotic use, oral hygiene). These variables were incorporated into separate XGBoost models for each group, both of which achieved strong predictive performance (AUC-ROC > 0.87). SHAP interpretation confirmed the importance and directionality of selected features, aligning with clinical findings. The integration of sPLS-DA and XGBoost provided both robust prediction and insights into the pathogenesis of oral candidiasis, supporting risk-based monitoring and prevention strategies in oncology care.

REFERENCES

1. Akpan A, Morgan R. Oral candidiasis. Postgraduate medical journal. 2002;78(922):455-459.

2.Lalla RV, Latortue MC, Hong CH, Ariyawardana A, D'Amato-Palumbo S, et al. A systematic review of oral fungal infections in patients receiving cancer therapy. Supportive care in cancer. 2010;18(8):985-992.

3. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2024;74(3):229-263.

4. Chitapanarux I, Wongsrita S, Sripan P, Kongsupapsiri P, Phakoetsuk P, et al. An underestimated pitfall of oral candidiasis in head and neck cancer patients undergoing radiotherapy: an observation study. BMC oral health. 2021;21(1):353.

5. Diaz PI, Hong B-Y, Dupuy AK, Choquette L, Thompson A, et al. Integrated analysis of clinical and microbiome risk factors associated with the development of oral candidiasis during cancer chemotherapy. Journal of Fungi. 2019;5(2):49.

6. Mayer LM, Strich JR, Kadri SS, Lionakis MS, Evans NG, et al., editors. Machine learning in infectious disease for risk factor identification and hypothesis generation: Proof of concept using invasive candidiasis. Proceedings of the Open forum infectious diseases; 2022: Oxford University Press.

7. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 2020;56(9):455.

8. Châu NTM, Hòa PTN. Tỷ lệ nhiễm nấm Candida niêm mạc miệng và yếu tố liên quan ở bệnh nhân điều trị bệnh nội khoa tại Bệnh viện Trường Đại học Y-Dược Huế. Tạp chí Y Dược học - Trường Đại học Y Dược Huế. 2023;3(13):126-132.

9. Nguyen BV, Nguyen HH, Vo T-H, Le M-T, Tran-Nguyen V-K, et al. Prevalence and drug susceptibility of clinical Candida species in nasopharyngeal cancer patients in Vietnam. One Health. 2024;18:100659.

10. Pulito C, Cristaudo A, Porta CL, Zapperi S, Blandino G, et al. Oral mucositis: the hidden side of cancer therapy. Journal of experimental & clinical cancer research. 2020;39:1-15.

11. Epstein JB, Thariat J, Bensadoun RJ, Barasch A, Murphy BA, et al. Oral complications of cancer and cancer therapy: from cancer treatment to survivorship. CA: a cancer journal for clinicians. 2012;62(6):400-422.

12. Patil S, Rao RS, Majumdar B, Anil S. Clinical appearance of oral Candida infection and therapeutic strategies. Frontiers in microbiology. 2015;6:1391.

13. Fidel Jr P. Candida-host interactions in HIV disease: implications for oropharyngeal candidiasis. Advances in dental research. 2011;23(1):45-49.

14. Dinarello CA. Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. Blood, The Journal of the American Society of Hematology. 2011;117(14):3720-3732.

15. Kubo M. Mast cells and basophils in allergic inflammation. Current opinion in immunology. 2018;54:74-79.

16. Tonzetich J. Production and origin of oral malodor: a review of mechanisms and methods of analysis. Journal of periodontology. 1977;48(1):13-20.

17. Li Z, Li J, Fu R, Liu Ja, Wen X, Zhang L. Halitosis: etiology, prevention, and the role of microbiota. Clinical oral investigations. 2023;27(11):6383-6393.

18. Coronado-Castellote L, Jiménez-Soriano Y. Clinical and microbiological diagnosis of oral candidiasis. Journal of clinical and experimental dentistry. 2013;5(5):e279.

19. Lu S-Y. Perception of iron deficiency from oral mucosa alterations that show a high prevalence of Candida infection. Journal of the Formosan Medical Association. 2016;115(8):619-627.

20. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930.

21. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-1358.

22. Shi J, Chen L, Yuan X, Yang J, Xu Y, et al. A potential XGBoost Diagnostic Score for Staphylococcus aureus bloodstream infection. Frontiers in Immunology. 2025;16:1574003.

23. Gu Y, Su S, Wang X, Mao J, Ni X, et al. Comparative study of XGBoost and logistic regression for predicting sarcopenia in postsurgical gastric cancer patients. Scientific Reports. 2025;15(1):12808.

24. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.