

Explainable Machine Learning for Predicting Skin Cancer Invasiveness: Insights into Regional Heterogeneity in Brazil

Maria Gonzales^{1*}, Jose Santos¹, Anna Cruz², Carlos Reyes³

¹ Department of Dermatology and Explainable AI, Faculty of Medicine, University of the Philippines Manila, Manila, Philippines.

² Department of Skin Cancer Epidemiology and Regional Health Disparities, Faculty of Medicine, University of Santo Tomas, Manila, Philippines.

³ Department of Invasiveness Prediction and Machine Learning, Faculty of Medicine, Ateneo de Manila University, Quezon City, Philippines.

*E-mail ✉ maria.gonzales@upm.edu.ph

Abstract

Skin cancer constitutes the most frequent neoplasm in Brazil, and its potential for invasion is shaped by several elements, among which geographical factors play a notable role. Recognizing these elements is essential for advancing diagnostic accuracy and therapeutic effectiveness. This investigation examined how geographical region influences skin cancer invasiveness in Brazil by identifying distinct predictive patterns associated with different regions. Data obtained from the Hospital Cancer Registries (RHC) maintained by Brazil's National Cancer Institute (INCA) underwent thorough analysis and preprocessing. Machine learning algorithms were then applied, and model interpretations were generated using the SHapley Additive exPlanations (SHAP) technique. The analysis indicated that geographical location is a key determinant in forecasting skin cancer invasiveness, underscoring the need to incorporate regional factors in future research. This study established that regional characteristics within Brazil significantly affect predictions of skin cancer invasiveness. While challenges such as data imbalance were encountered, the outcomes provide valuable insights for crafting more targeted policies to combat skin cancer across the country.

Keywords: Skin cancer, Machine learning, SHAP approach, Comparison between regions, Cancer invasive behavior

Introduction

Skin cancer involves the uncontrolled growth of cutaneous cells [1]. In 2021, the condition recorded the highest incidence rates both nationally in Brazil and globally [2]. Several subtypes exist, each determined by the particular skin layer affected [3]. As shown in **Figure 1**, melanoma occurs less commonly than other forms yet is the most aggressive and responsible for most deaths [4], primarily because of its strong tendency to invade

distant tissues [5]. In skin cancer progression, invasion starts at the primary lesion, where malignant cells breach the basement membrane, enter the dermis, and enter the circulatory system [6], potentially enabling spread to distant anatomical sites [7].

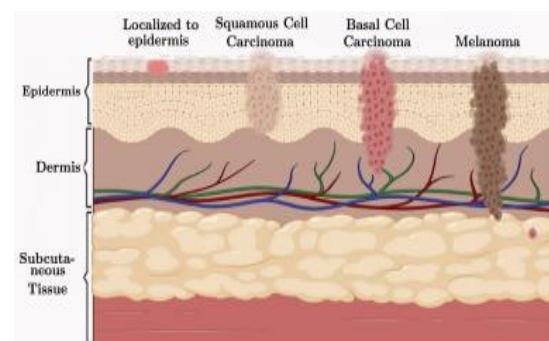


Figure 1. Types of skin cancer [8].

Access this article online

<https://smerpub.com/>

Received: 09 September 2025; **Accepted:** 07 December 2025

Copyright CC BY-NC-SA 4.0

How to cite this article: Gonzales M, Santos J, Cruz A, Reyes C. Explainable Machine Learning for Predicting Skin Cancer Invasiveness: Insights into Regional Heterogeneity in Brazil. J Med Sci Interdiscip Res. 2025;5(2):163-76. <https://doi.org/10.51847/5v9Z0jqCoE>

Consequently, invasive tumors tend to produce more intense pain resulting from compression of adjacent structures and the inflammatory processes elicited by tissue penetration [9]. Accurate differentiation between invasive and non-invasive lesions is therefore critical for determining prognosis and safeguarding patient well-being during therapy. In Brazil, skin cancer accounts for 27% of all diagnosed malignancies, with projections indicating approximately 220,490 new cases each year from 2023 to 2025 [10].

The substantial burden imposed by skin cancer and its management costs prompted Brazil's Unified Health System (SUS) to increase chemotherapy expenditures by roughly 47 million Brazilian reais in 2022 [11]. However, the allocation of healthcare infrastructure and resources across the country's administrative regions continues to show marked inequalities [12]. Such disparities in service availability strongly influence diagnostic timeliness and treatment success, with the South benefiting from superior access, the North facing the greatest limitations, and the Northeast, Southeast, and Midwest presenting intermediate conditions [13].

In addition, contemporary technological developments have led to wider adoption of machine learning techniques in diverse domains, with particularly promising applications in cancer research [14]. These approaches excel at handling complex datasets, performing feature selection, and generating reliable predictive outcomes [15]. Model interpretability is equally vital for meaningful clinical application [16]. The SHAP framework enables the development of transparent models that reveal the relative importance of individual features in driving machine learning predictions [17].

Against this background, the current research analyzes and contrasts Brazil's regions to assess their effect on skin cancer invasiveness. The study addresses three core dimensions: determining the predictive relevance of geographical region, identifying region-specific features

linked to invasiveness, and comparing regions based on the variables that define invasive potential. These variables underpin the observed predictive patterns. Clarifying these regional differences is fundamental to better understanding skin cancer dynamics and informing tailored prevention and intervention strategies.

The urgency of this line of inquiry is reinforced by the disease's elevated incidence both domestically and worldwide, coupled with substantial regional inequities in healthcare access within Brazil. Furthermore, accurate, interpretable machine learning outputs can meaningfully strengthen skin cancer prevention and treatment initiatives [18].

Materials and Methods

The dataset employed in this investigation was obtained from the Hospital Cancer Registry (RHC) maintained by the National Cancer Institute (INCA). It comprises 4,680,029 anonymized records of patients diagnosed with cancer and includes 47 variables.

Initial data preparation

The database underwent a comprehensive examination to characterize its structure and properties. Exploratory data analysis was performed to identify relationships among variables. The original files were converted from DBF to CSV format, after which a systematic filtering procedure was applied. Additionally, new variables were derived from the existing ones to align with the study's specific objectives.

Selection of samples and variables

The initial step involved refining the sample set through a systematic selection process, as outlined in the flowchart shown in **Figure 2**. This procedure reduced the dataset to 48,331 records. The final sample included cases from all five Brazilian regions, as depicted in **Figure 3** and summarized in **Table 1**.

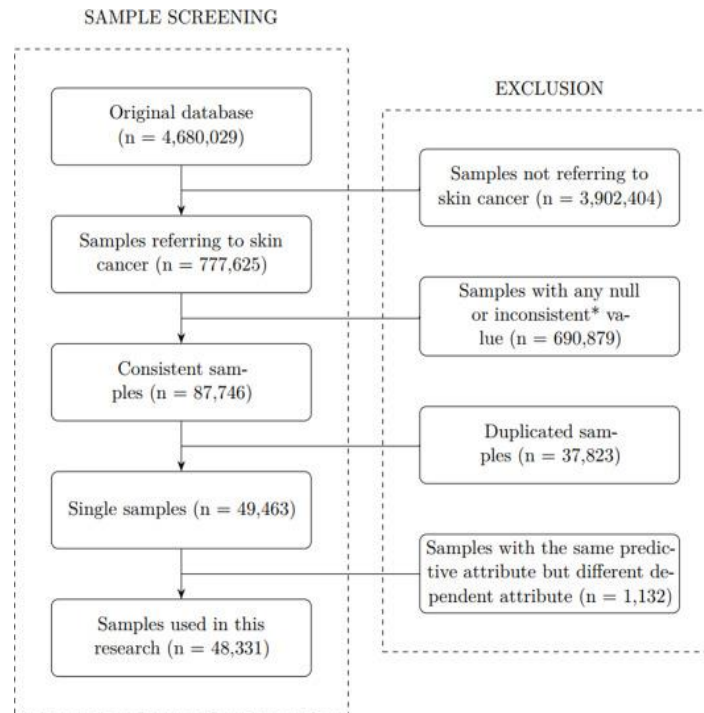


Figure 2. Screening flowchart of the samples

* Inconsistent entries, such as the value '3' in the gender field (where '1' denotes male and '2' denotes female according to the data dictionary), were identified as invalid and resulted in the removal of the affected records (Source: Compiled by the author with data extracted from the research).

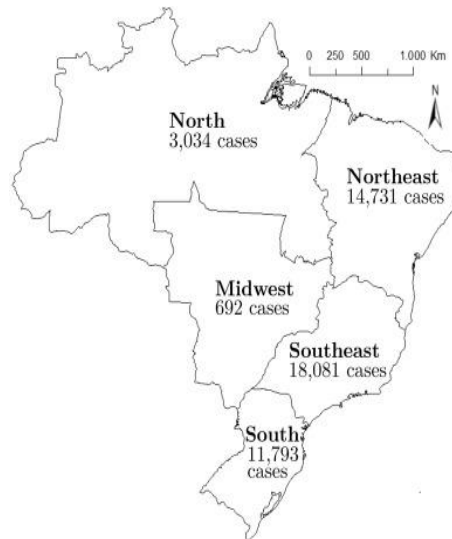


Figure 3. Distribution of cases in Brazil (Source: Compiled by the author with data extracted from the research).

Table 1. Distribution of the dependent variable according to Brazilian region (Source: Compiled by the author with data extracted from the research).

Region	Invasive	Non-Invasive	% Invasives	Population ²
Midwest	661	31	95.52 %	16,289,538
North	3,004	30	99.01 %	17,354,884
Northeast	14,494	237	98.39 %	54,658,515

Southeast	17,408	673	96.28 %	84,840,113
South	11,416	377	96.80 %	29,937,706
Brazil (Total)	46,983	1,348	97.21 %	203,080,756

^a Population figures are based on the 2022 census conducted by the Brazilian Institute of Geography and Statistics (IBGE) [19].

Variable selection was carried out to improve data quality by eliminating attributes that showed no variation, lacked relevance to the predictive task, or were redundant because they depended on other variables. Following data encoding, highly correlated variables were removed using Cramér's V coefficient, retaining only one representative from each strongly associated pair.

Table 2 presents the complete list of variables included in the analysis, along with their respective characteristics. For each variable, the table describes its domain (possible

values or categories) and assigns numerical codes to each category to facilitate processing by machine learning algorithms. The dependent variable in this study is cancer invasiveness, classified as either invasive or non-invasive. Its distribution across Brazilian regions is displayed in **Table 1**. The dataset showed notable class imbalance, with invasive cancers as the majority class of interest and non-invasive cases as the minority class, as indicated in the same table.

Table 2. Attributes used in research (Source: Compiled by the author with data extracted from the research).

Variable	Categories and encoded values
Primary tumor site	Eyelid (0); External ear (1); Face (2); Scalp and neck (3); Trunk (4); Shoulders and upper extremities (5); Hips and lower extremities (6); Overlapping skin lesion (7)
Region	Midwest (0); North (1); Northeast (2); Southeast (3); South (4)
Age category	000–009 (0); 010–019 (1); 020–029 (2); 030–039 (3); 040–049 (4); 050–059 (5); 060–069 (6); 070–079 (7); 080–089 (8); 090–103 (9)
Marital status	Single (0); Married (1); Widowed (2); Separated (3); Common-law union (4)
Educational attainment	No formal education (0); Incomplete elementary education (1); Completed elementary education (2); High school education (3); Incomplete undergraduate degree (4); Completed undergraduate degree (5)
Ethnicity/Race	White (0); Black (1); Yellow (2); Brown (3); Indigenous (4)
History of alcohol use	Never consumed (0); Former consumer (1); Current consumer (2)
History of tobacco use	Never smoked (0); Former smoker (1); Current smoker (2)
Referral source ^a	SUS (0); Non-SUS (1); Self-referred (2)
Case classification ^b	Analytic (0); Non-analytic (1)
Family history of cancer	Yes (0); No (1)
Multiple primary tumors	Yes (0); No (1)
Sex	Male (0); Female (1)
Cancer invasiveness (target variable)	Non-invasive (0); Invasive (1)

^a In the context of cancer care, this variable indicates the source of patient referral. It distinguishes referrals from the public Unified Health System (SUS) (code 1), private services or health insurance plans (code 2), self-referrals (code 3), or follow-up cases of previously diagnosed cancers (code 8).

^b This variable categorizes cases as “analytic” when both the initial diagnosis and treatment occur at the registering hospital of INCA, or “non-analytic” when diagnosis or treatment took place elsewhere or deviated from the registering hospital's protocols. Analytic cases receive detailed monitoring and evaluation, whereas non-analytic cases are registered with limited follow-up.

Selection of machine learning algorithms

This study utilized the following machine learning algorithms: XGBoost (XGB), Support-Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Naïve Bayes (NB), and Decision Tree (DT). The selection was informed by earlier investigations from Ghazal *et al.* [20] and Liu *et al.* [21], which established the reliability and

relevance of these algorithms in similar analytical settings.

Following the procedure outlined by Yu *et al.* [22], the dataset was randomly divided into separate training and testing portions. Eighty percent (80%) of the records were assigned to the training set for model construction, with the remaining twenty percent (20%) reserved for performance evaluation on unseen data.

We then optimized hyperparameters on the training data using grid search with ten-fold cross-validation, as proposed by the same authors. This strategy helped determine optimal parameter settings while accounting for class imbalance. We selected the parameter combination that delivered the best balanced accuracy. We conducted model development and tuning both nationally and by region for each of Brazil's five geographical regions, accounting for potential regional differences.

To address the existing class imbalance and ensure robust cross-validation, we used balanced accuracy as the primary performance indicator, in accordance with recommendations by Luo *et al.* [23]. This metric is particularly well-suited for assessing classification quality in datasets with unequal class distributions. We did not implement methods such as Synthetic Minority Over-sampling Technique (SMOTE), threshold modification, or algorithmic weighting. The emphasis remained on examining the algorithms' native performance using the unaltered dataset to maintain its original integrity.

Supplementary evaluation relied on three additional indicators: sensitivity, to gauge the detection rate of positive cases; specificity, to measure the correct classification of negative cases; and the Matthews Correlation Coefficient (MCC), which provides a comprehensive evaluation for binary outcomes. This combination of metrics aligns with the approach adopted by Sasikala, Ezhilarasi and Arunkumar [24].

Identification of the most effective algorithms for predicting cancer invasiveness

Algorithm performance proved largely consistent across the different Brazilian regions and aligned closely with national-level outcomes. Among all tested models, the Support-Vector Machine (SVM) delivered the highest results both regionally and at the country-wide scale. **Figure 4** highlights SVM's leading position in terms of MCC and balanced accuracy, reflecting its strong ability to limit misclassifications, a critical requirement in medical decision-making contexts [25]. The national test set confusion matrix in **3** further confirms its predictive strength.

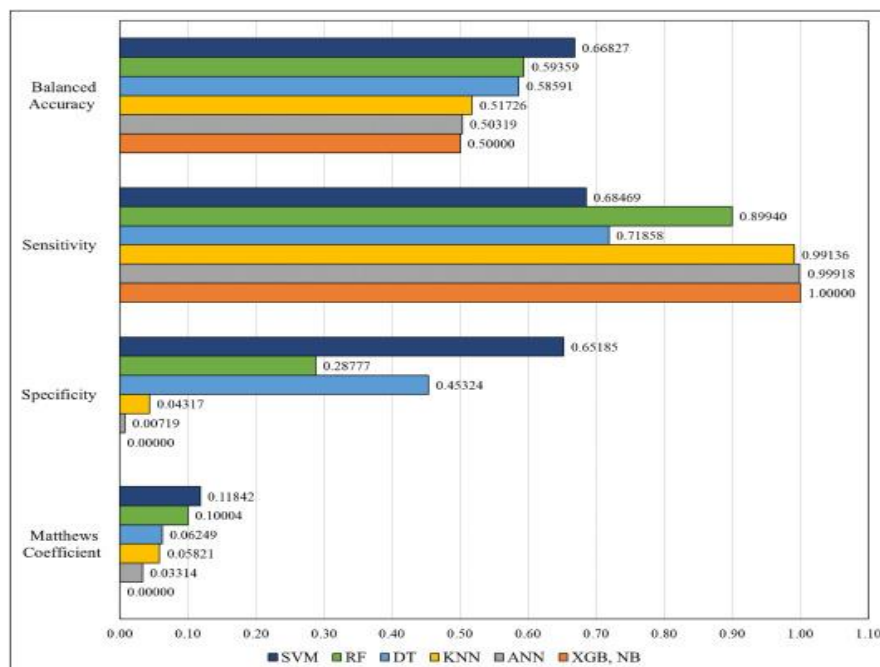


Figure 4. Ranking of models on a national scope (Source: Compiled by the author with data extracted from the research).

Table 3. Binary confusion matrix on a national scope (Source: Compiled by the author with data extracted from the research).

Actual \ Predicted	Invasive	Non-Invasive
Invasive	6434	2963

Non-Invasive

94

176

Development of explainability models using SHAP

Explainability models serve to reveal the variables that exert the greatest influence on the outputs of machine learning algorithms [26]. Among available interpretability techniques, the SHAP method was adopted, following the framework introduced by Lundberg and Lee [27] and applied in related studies by Moncada-Torres *et al.* [28] and Huang *et al.* [29]. Understanding the SHAP approach requires familiarity with Shapley values. These values represent the weighted average of an element's marginal contributions across all possible coalitions in a cooperative game, thereby quantifying its overall contribution to the outcome [30]. The SHAP framework produces local SHAP values that quantify each variable's contribution to the model's prediction for an individual instance in the dataset [27]. These local values may be positive or negative, signifying that a given variable can either increase or decrease the likelihood of a particular outcome for the dependent variable [31]. Global SHAP values are derived from the local values and reflect each variable's overall influence across the complete dataset [32]. They are computed as the mean of the absolute local SHAP values for that variable over all observations [31].

Once the optimal algorithm and corresponding hyperparameters were identified for both national and regional levels, the subsequent steps focused on model

training, evaluation, and interpretation. The selected algorithm, SVM, was fitted to the training data for each regional dataset and the national dataset. To identify the most influential variables driving the predictions, we implemented the SHAP method. We chose KernelExplainer because it is compatible with SVM models. This explainer assesses model behavior by generating predictions across varied feature combinations through sampling and computes the average contribution of each feature [33]. We then divided the local SHAP values into two groups corresponding to contributions toward non-invasive and invasive cancer classifications.

Results and Discussion

Figure 5 presents the global SHAP values arranged along the horizontal axis, with variables ordered vertically from highest to lowest importance. Each horizontal bar illustrates the degree of influence of the respective variable, where bar length corresponds to the strength of its effect. The exact global SHAP value for each variable appears at the end of its bar. At the national level, the analysis identified the primary tumor site as the most influential variable, followed by geographical region, presence of multiple tumors, educational attainment, and sex.

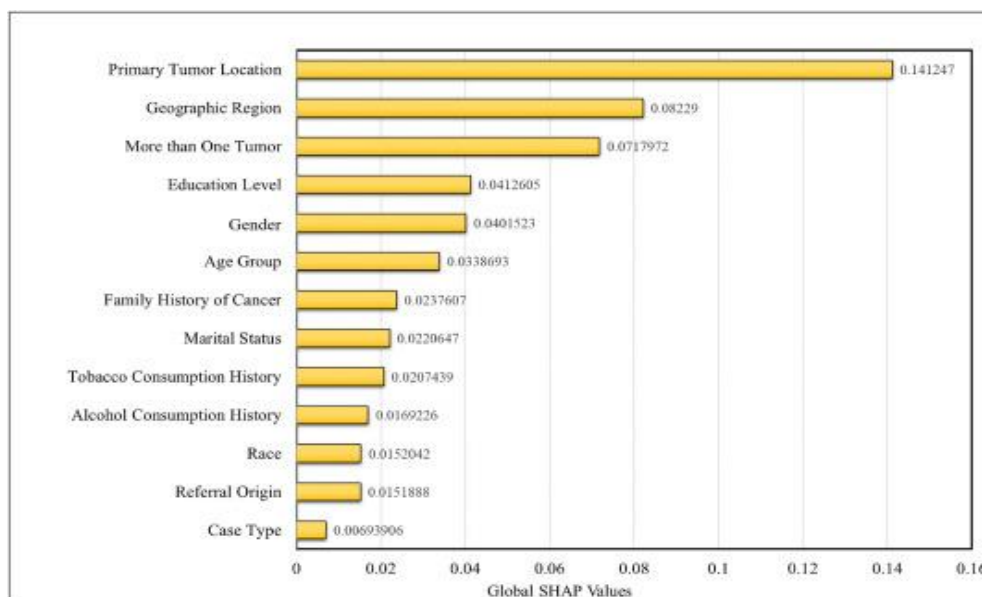


Figure 5. Impact of the attributes at the national level (Source: Compiled by the author with data extracted from the research).

Similar to **Figure 5** and **Figure 6** illustrates the global SHAP values for every attribute across all Brazilian regions. In this visualization, each bar is segmented into distinct colors representing the five regions. The length

and placement of these colored segments reflect each variable's relative importance within the predictive models for each region, with longer segments indicating stronger influence.

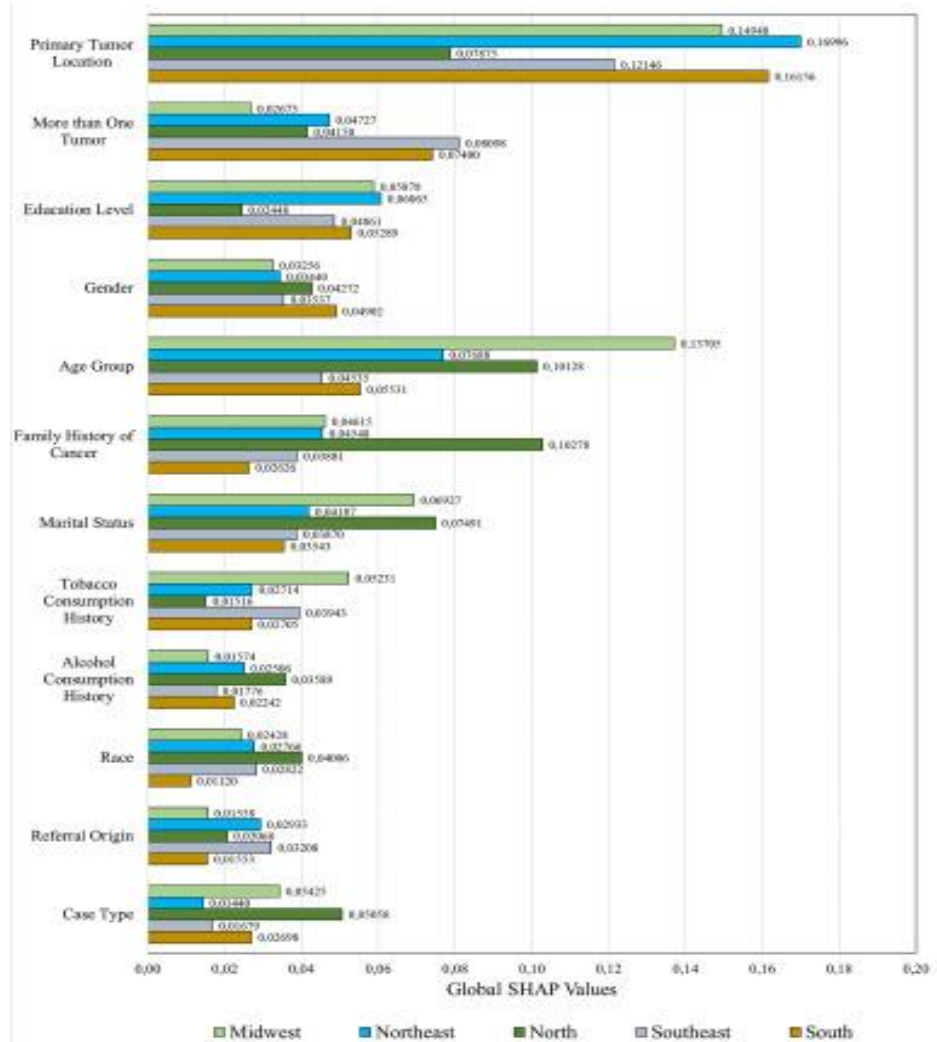


Figure 6. Importance of attributes by region.

Figure 7 illustrates how individual variables affect model predictions for cancer invasiveness, ranked by importance. Variables appear on the vertical axis, color-coded according to the categories defined in **Table 2** (blue representing lower numerical codes and red higher codes). The horizontal axis displays local SHAP values, where positive values contribute toward an invasive

classification and negative values toward a non-invasive classification. Point density indicates how frequently particular local SHAP values occur for each variable. In contrast, the distance of points from the central vertical line reflects the magnitude of their contribution to the final prediction.

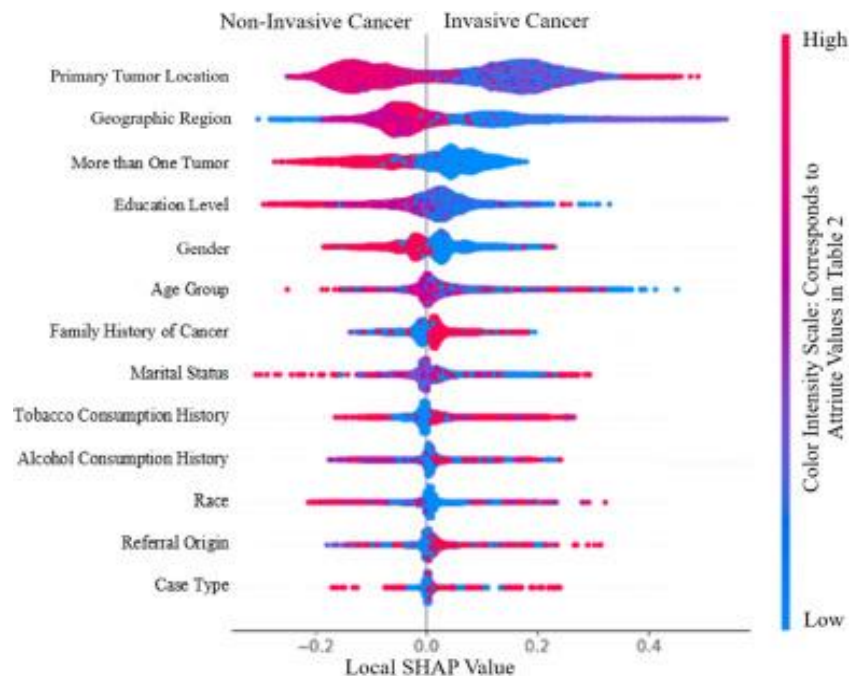


Figure 7. Attribute prediction pattern at the national level (Source: Compiled by the author with data extracted from the research).

This investigation evaluated the influence of regional variables on predicting skin cancer invasiveness using explainable artificial intelligence, specifically the SHAP framework. The findings presented in Section 3 were examined against the three central objectives established in Section 1: evaluating the contribution of geographical region to predicting invasive versus non-invasive status, identifying region-specific variables associated with cancer invasiveness, and comparing regions based on the characteristics that shape invasive behavior.

Influence of geographical region on predicting cancer invasiveness

The assessment of the regional variable's importance, as illustrated in **Figure 5**, shows that the primary tumor site is the dominant factor, with a markedly higher global SHAP value than all other variables. This indicates that

the tumor's anatomical origin plays a decisive role in determining its invasive potential. Although geographical region ranks as the second most influential variable, its global SHAP value is substantially lower, suggesting that while it contributes meaningfully to the prediction, its effect is secondary to that of tumor location.

Figure 8 further supports this observation by showing mean local SHAP values by anatomical tumor site, stratified by geographical region and for Brazil overall. These results emphasize the effect of tumor location on invasiveness. Unlike **Figure 5** and **Figure 8** does not depict global SHAP values, since the presented averages are not absolute. Negative values in this figure correspond to contributions toward non-invasiveness, whereas positive values indicate contributions toward invasiveness.

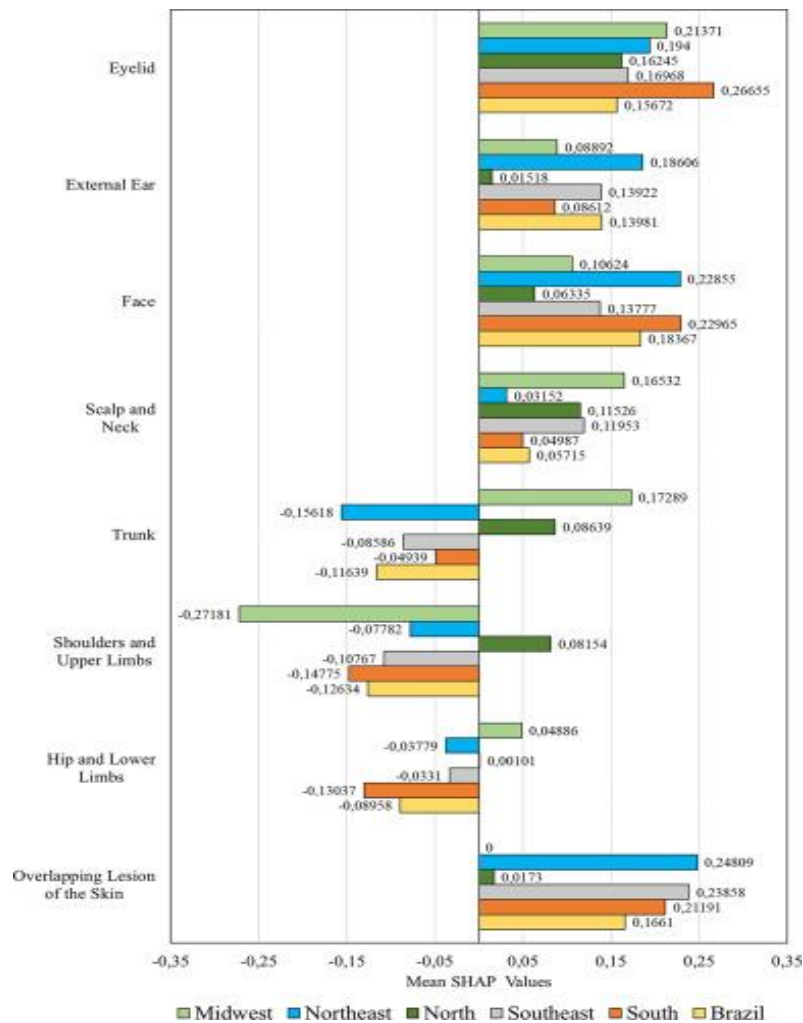


Figure 8. Impact of anatomic attributes by region.

Despite its secondary position, the geographical region variable possesses a considerable global SHAP value that exceeds those of several other factors, including history of tobacco and alcohol consumption, ethnicity, and source of referral. This finding suggests a moderate but relevant influence of geographical region on predicting cancer invasiveness. Such importance may stem from underlying elements not directly captured by the model, such as disparities in healthcare availability, environmental exposures, or region-specific lifestyle patterns. Therefore, although not the leading predictor, geographical region contributes substantially to the model's overall predictive capacity.

Detection of region-specific patterns in predictions

Examination of the global SHAP values in **Figure 6** uncovers notable differences among regions regarding the variables that influence predictions of skin cancer

invasiveness in Brazil. In the South, the occurrence of multiple tumors emerges as a particularly strong predictor, with age group and gender also demonstrating meaningful contributions. This pattern suggests that demographic characteristics are particularly relevant in this part of the country.

In the Southeast, the primary tumor location stands out as the overwhelmingly dominant factor, indicating that the anatomical site of origin is the main determinant of invasiveness. Although gender and age group exert some influence, their effects are comparatively modest, pointing to intricate interactions among biological variables in this region. The North presents a different profile, where age group exerts the strongest effect, possibly reflecting distinct demographic structures or localized risk exposures. Gender also plays a role, albeit to a lesser extent, highlighting potential gender-related differences in disease behavior.

In the Northeast, both primary tumor location and age group predominate, underscoring the combined importance of anatomical site and patient age in forecasting invasiveness. Gender also contributes, which may reflect differences in health-seeking behaviors and access to care between men and women in the region. In the Midwest, primary tumor location again shows the greatest impact, closely followed by age group. Together, these variables highlight the prominent roles of cancer type and patient age in determining invasiveness within this area.

Overall, the regional analysis demonstrates that primary tumor location consistently ranks among the most influential predictors of skin cancer invasiveness across all regions of Brazil. Age group likewise emerges as a key factor in nearly every region, emphasizing the relevance of patient age in disease prognosis. Gender,

despite variations in strength, remains significant nationwide. These differences illustrate the diverse and multifaceted factors driving skin cancer invasiveness across the country.

Comparison of regional characteristics influencing cancer invasiveness

In conjunction with the visualization in **Figure 7**, **Table 4** and **Table 5** provide further clarity on the predictive profiles associated with invasive and non-invasive skin cancer throughout Brazil's geographical regions. The tables present the ten most influential characteristics for each region, arranged by the country's five major areas. Individual entries specify particular categories within attribute domains that are associated with increased risk of invasive disease.

Table 4. Patterns of invasive skin cancer prediction by region (Source: Compiled by the author with data extracted from the research).

Rank	Midwest	Northeast	North	Southeast	South
1st	Eyelid (<i>Primary Tumor Site</i>)	030–039 (<i>Age Category</i>)	000–009 (<i>Age Category</i>)	Overlapping skin lesion (<i>Primary Tumor Site</i>)	Eyelid (<i>Primary Tumor Site</i>)
2nd	070–079 (<i>Age Category</i>)	Overlapping skin lesion (<i>Primary Tumor Site</i>)	Non-SUS (<i>Referral Source</i>)	Eyelid (<i>Primary Tumor Site</i>)	Face (<i>Primary Tumor Site</i>)
3rd	Common-law union (<i>Marital Status</i>)	Face (<i>Primary Tumor Site</i>)	Common-law union (<i>Marital Status</i>)	External ear (<i>Primary Tumor Site</i>)	Overlapping skin lesion (<i>Primary Tumor Site</i>)
4th	Incomplete undergraduate degree (<i>Educational Attainment</i>)	010–019 (<i>Age Category</i>)	Self-referred (<i>Referral Source</i>)	Face (<i>Primary Tumor Site</i>)	No formal education (<i>Educational Attainment</i>)
5th	Completed undergraduate degree (<i>Educational Attainment</i>)	020–029 (<i>Age Category</i>)	040–049 (<i>Age Category</i>)	Scalp and neck (<i>Primary Tumor Site</i>)	010–019 (<i>Age Category</i>)
6th	Trunk (<i>Primary Tumor Site</i>)	Eyelid (<i>Primary Tumor Site</i>)	020–029 (<i>Age Category</i>)	090–103 (<i>Age Category</i>)	020–029 (<i>Age Category</i>)
7th	Scalp and neck (<i>Primary Tumor Site</i>)	External ear (<i>Primary Tumor Site</i>)	Separated (<i>Marital Status</i>)	Common-law union (<i>Marital Status</i>)	030–039 (<i>Age Category</i>)
8th	Current consumer (<i>Tobacco Consumption History</i>)	000–009 (<i>Age Category</i>)	010–019 (<i>Age Category</i>)	Current consumer (<i>Tobacco Consumption History</i>)	External ear (<i>Primary Tumor Site</i>)
9th	Separated (<i>Marital Status</i>)	No formal education (<i>Educational Attainment</i>)	No (<i>More than One Tumor</i>)	Separated (<i>Marital Status</i>)	Single (<i>Marital Status</i>)
10th	080–089 (<i>Age Category</i>)	040–049 (<i>Age Category</i>)	Eyelid (<i>Primary Tumor Site</i>)	Yellow (<i>Race</i>)	090–103 (<i>Age Category</i>)

Table 5. Patterns of non-invasive skin cancer prediction by region (Source: Compiled by the author with data extracted from the research).

Rank	Midwest	Northeast	North	Southeast	South
1st	Shoulders and upper extremities (<i>Primary Tumor Site</i>)	Trunk (<i>Primary Tumor Site</i>)	Yes (<i>Family History of Cancer</i>)	Shoulders and upper extremities (<i>Primary Tumor Site</i>)	Shoulders and upper extremities (<i>Primary Tumor Site</i>)
2nd	000–009 (<i>Age Category</i>)	Shoulders and upper extremities (<i>Primary Tumor Site</i>)	070–079 (<i>Age Category</i>)	000–009 (<i>Age Category</i>)	Hips and lower extremities (<i>Primary Tumor Site</i>)
3rd	030–039 (<i>Age Category</i>)	No (<i>More than One Tumor</i>)	Yes (<i>History of Alcohol Use</i>)	No (<i>More than One Tumor</i>)	No (<i>More than One Tumor</i>)
4th	010–019 (<i>Age Category</i>)	Hips and lower extremities (<i>Primary Tumor Site</i>)	Married (<i>Marital Status</i>)	Trunk (<i>Primary Tumor Site</i>)	Trunk (<i>Primary Tumor Site</i>)
5th	040–049 (<i>Age Category</i>)	060–069 (<i>Age Category</i>)	Completed elementary education (<i>Educational Attainment</i>)	010–019 (<i>Age Category</i>)	Female (<i>Sex</i>)
6th	020–029 (<i>Age Category</i>)	090–103 (<i>Age Category</i>)	Yes (<i>More than One Tumor</i>)	Incomplete undergraduate degree (<i>Educational Attainment</i>)	Indigenous (<i>Race</i>)
7th	Married (<i>Marital Status</i>)	Female (<i>Sex</i>)	SUS (<i>Referral Source</i>)	Completed undergraduate degree (<i>Educational Attainment</i>)	Black (<i>Race</i>)
8th	White (<i>Race</i>)	Completed undergraduate degree (<i>Educational Attainment</i>)	Former smoker (<i>History of Tobacco Use</i>)	Hips and lower extremities (<i>Primary Tumor Site</i>)	Brown (<i>Race</i>)
9th	Yes (<i>Family History of Cancer</i>)	Married (<i>Marital Status</i>)	Never smoked (<i>History of Tobacco Use</i>)	Yes (<i>Family History of Cancer</i>)	Analytic (<i>Case Classification</i>)
10th	Analytic (<i>Case Classification</i>)	Yes (<i>Family History of Cancer</i>)	Incomplete elementary education (<i>Educational Attainment</i>)	Non-SUS (<i>Referral Source</i>)	Separated (<i>Marital Status</i>)

Predictive pattern for invasive skin cancer

Figure 7 reveals that the dominant pattern for invasive skin cancer occurs more frequently in cases where the primary tumor is situated in the neck or superior body regions. The color scale reinforces this observation, and the regional analysis in **Figure 8** corroborates it. The pattern is also more pronounced among patients receiving care in the North and Northeast regions of Brazil. Affected individuals span a wide age spectrum, ranging from 20 to 103 years, which suggests that age may interact with other variables rather than acting as a standalone determinant.

Other contributing factors, though less dominant, also contribute to this predictive profile. These include lower educational levels, potentially linked to reduced health literacy, and marital status, with higher frequencies of legal separation or informal unions.

Table 4 details the leading characteristics associated with invasive skin cancer predictions across the Midwest, Northeast, North, Southeast, and South. Regional differences are evident in the priority given to specific risk and demographic elements. For instance, in the Midwest, the strongest indicators include primary tumor on the eyelid, age group 070–079, and consensual union.

In the Northeast, prominent features are age group 030–039, overlapping skin lesions, and facial tumor location. The North highlights referral from non-SUS sources, consensual union, and age group 000–009. In the Southeast, face, overlapping lesions, and external ear locations predominate. In the South, eyelid and facial tumor sites, age group 010–019, and completion of a bachelor's degree stand out. Variables such as age group, marital status, and tobacco consumption history appear consistently across several regions, indicating shared national risk tendencies.

Predictive pattern for non-invasive skin cancer

Figure 7 further identifies a characteristic profile for non-invasive skin cancer, which is more common in female patients with tumors below the neck, as also indicated in **Figure 8**. This pattern occurs more frequently in the South and Southeast regions. Again, a broad age distribution is observed, implying multifaceted interactions involving age and additional variables.

Table 5 outlines the combination of demographic and clinical features linked to non-invasive predictions in each region. In the Midwest, notable factors include primary tumor sites on the shoulders and upper limbs, age group 000–009, married status, and white ethnicity. The Northeast features tumor locations on the trunk, shoulders, and upper limbs; female sex; age groups 060–069 and 090–103; and bachelor's degree attainment. In the North, relevant elements are family cancer history, age group 070–079, married status, elementary education completion, former tobacco use, and SUS referral. For the Southeast, important characteristics encompass shoulder and upper limb tumor sites, age groups 000–009 and 010–109, absence of multiple tumors, incomplete bachelor's degree, white ethnicity, and family cancer history. In the South, the profile includes tumor locations on shoulders, upper limbs, hips, and lower limbs, absence of multiple tumors, female sex, black or brown ethnicity, and separated marital status.

Conclusion

This research emphasizes the critical role of geographical variations within Brazil when predicting skin cancer invasiveness. By integrating predictive modeling with SHAP-based interpretations, the study demonstrates that region of residence not only directly influences the likelihood of invasive disease but also modifies the weight of other predictive variables, resulting in distinct

regional signatures. Such findings stress the importance of developing customized, region-sensitive strategies to strengthen prevention and treatment outcomes for skin cancer.

Validity threats

Regarding validity concerns, the dataset shows a marked predominance of invasive over non-invasive cases. This imbalance poses an internal validity risk through possible model bias toward the majority class. Externally, the distribution may not mirror true population prevalence, which could constrain the applicability of results in other settings. Mitigation efforts, detailed in subsection 2.3, focused on selecting algorithms based on balanced accuracy performance.

Future Research

Subsequent investigations are encouraged to extend the current methodological framework to other commonly occurring cancers in Brazil, including breast and prostate cancer, to elucidate additional region-specific patterns and determinants.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Skin cancer [Internet]. Encyclopedia Britannica. 2023 [cited 2026 Aug 23]. Available from: <https://www.britannica.com/science/skin-cancer>
2. Câncer de pele [Internet]. Brazilian Ministry of Health [cited 2026 Aug 23]. Available from: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-pele>
3. American Cancer Society. What Are Basal and Squamous Cell Skin Cancers? [Internet]. Cancer.org. 2019 [cited 2026 Aug 23]. Available from: <https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/what-is-basal-and-squamous-cell.htm>

4. Davey MG, Miller N, McInerney NM. A review of epidemiology and cancer biology of Malignant Melanoma. *Cureus*. 2021;13(5):e15087. doi:10.7759/cureus.15087
5. National Cancer Institute. Skin cancer (including melanoma)—patient version [Internet]. Cancer.gov. 2019 [cited 2026 Aug 23]. Available from: <https://www.cancer.gov/types/skin>
6. National Cancer Institute. Genetics of Skin Cancer (PDQ®)—Health Professional Version [Internet]. Cancer.gov. 2019 [cited 2026 Aug 23]. Available from: <https://www.cancer.gov/types/skin/hp/skin-genetics-pdq>
7. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*. 2007;449(7163):682-8. doi:10.1038/nature06174
8. Hasan N, Nadaf A, Imran M, Jiba U, Sheikh A, Almalki WH, et al. Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches. *Mol Cancer*. 2023;22(1):168. doi:10.1186/s12943-023-01860-1
9. Fidler IJ. The pathogenesis of cancer metastasis: the “seed and soil” hypothesis revisited. *Nat Rev Cancer*. 2003;3(6):453-8. doi:10.1038/nrc1098
10. Santos MO, Lima FC, Martins LFL, Oliveira JFP, Almeida LM, Cancela MC. Estimativa de Incidência de Câncer no Brasil, 2023-2025. *Rev Bras Cancerol*. 2023;69(1):e-213458. doi:10.32635/2176-9745.RBC.2023v69n1.3458
11. Governo Federal reajusta em R\$ 47 milhões os valores para tratamento quimioterápico do câncer de pele [Internet]. Brazilian Ministry of Health. 2022 Mar [cited 2026 Aug 23]. Available from: <https://www.gov.br/saude/pt-br/assuntos/noticias/2022/marco/governo-federal-reajusta-em-r-47-milhoes-os-valores-para-tratamento-quimioterapico-do-cancer-de-pele>
12. Viacava F, Bellido JG. Condições de saúde, acesso a serviços e fontes de pagamento, segundo inquéritos domiciliares. *Cien Saude Colet*. 2016;21(2):351-70. doi:10.1590/1413-81232015212.20292015
13. Szwarcwald CL, Souza Júnior PRB, Marques AP, Almeida WS, Montilla DER. Inequalities in healthy life expectancy by Brazilian geographic regions: findings from the National Health Survey, 2013. *Int J Equity Health*. 2016;15(1):184. doi:10.1186/s12939-016-0471-7
14. Subasi A, Panigrahi SS, Patil BS, Canbaz MA, Klén R. Advanced pattern recognition tools for disease diagnosis. In: *5G IoT and Edge Computing for Smart Healthcare*. Elsevier; 2022. p. 195-229. doi:10.1016/B978-0-323-91143-2.00010-1
15. Machado F. *Big Data O Futuro dos Dados e Aplicações*. 1st ed. São Paulo: Editora Érica; 2018.
16. Molnar C. *Interpretable Machine Learning*. 2022 [cited 2026 Aug 23]. Available from: <https://christophm.github.io/interpretable-ml-book/>
17. Silva W, Oliveira V, Araújo S, Vieira D, Castro M. Explainability e auditability: interpretando e validando modelos de machine learning. In: *Advances in Artificial Intelligence and Data Science*. 2022. p. 55-78. doi:10.1016/B978-0-12-824536-1.00004-6
18. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017 Jan 25;542(7639):115-8. doi:10.1038/nature21056
19. IBGE. List of Municipal Populations to be published in the Official Gazette in 2023 [Internet]. 2023 [cited 2024 Feb 9]. Available from: <https://www.ibge.gov.br/en/statistics/social/population/37757-relacao-da-populacao-dos-municipios-para-publicacao-no-dou-2.html>
20. Ghazal TM, Al Hamadi H, Umar Nasir M, Atta-ur-Rahman, Gollapalli M, Zubair M, et al. Supervised Machine Learning Empowered Multifactorial Genetic Inheritance Disorder Prediction. *Comput Intell Neurosci*. 2022;2022:9366238. doi:10.1155/2022/9366238
21. Liu WC, Li MX, Qian WX, Luo ZW, Liao WJ, Liu ZL, et al. Application of Machine Learning Techniques to Predict Bone Metastasis in Patients with Prostate Cancer. *Cancer Manag Res*. 2021;13:8723-36. doi:10.2147/CMAR.S327581
22. Yu J, Zhou Y, Yang Q, Liu X, Huang L, Yu P, et al. Machine learning models for screening carotid atherosclerosis in asymptomatic adults. *Sci Rep*. 2021;11(1):22336. doi:10.1038/s41598-021-01869-0
23. Luo Y, Tseng HH, Cui S, Wei L, Ten Haken RK, El Naqa I. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR Open*. 2019;1(1):20190021. doi:10.1259/bjro.20190021

24. Sasikala S, Ezhilarasi M, Arunkumar S. Feature selection algorithm based on binary BAT algorithm and optimum path forest classifier for breast cancer detection using both echographic and elastographic mode ultrasound images. *J Cancer Res Ther.* 2023 Jan 1;19(2):191-7. doi:10.4103/jcrt.jcrt_1721_21
25. Herington J, McCradden MD, Creel K, Boellaard R, Jones E, Jha AK, et al. Ethical considerations for artificial intelligence in medical imaging: deployment and governance. *J Nucl Med.* 2023;64(10):1509-15. doi:10.2967/jnumed.123.265897
26. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv [Internet].* 2017 Feb 28 [cited 2026 Aug 23]; Available from: <https://arxiv.org/abs/1702.08608>
27. Lundberg S, Lee S. A Unified Approach to Interpreting Model Predictions. *arXiv [Internet].* 2017 May 22 [cited 2026 Aug 23]; Available from: <https://arxiv.org/abs/1705.07874v2>
28. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep.* 2021;11(1):7024. doi:10.1038/s41598-021-86332-8
29. Huang T, Le D, Yuan L, Xu S, Peng X. Machine learning for prediction of in-hospital mortality in lung cancer patients admitted to intensive care unit. *PLoS One.* 2023;18(1):e0280606. doi:10.1371/journal.pone.0280606
30. Shapley LS. 17. A Value for n-Person Games. In: *Contributions to the Theory of Games (AM-28), Volume II.* Princeton University Press; 1953. p. 307-18. doi:10.1515/9781400881970-018
31. Lundberg SM, Erion GG, Lee SI. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv [Internet].* 2018 Feb 8 [cited 2026 Aug 23]; Available from: <https://arxiv.org/abs/1802.03888>
32. Meng Y, Yang N, Qian Z, Zhang G. What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. *J Theor Appl Electron Commer Res.* 2020;16(3):466-90. doi:10.3390/jtaer16030028
33. SHAP KernelExplainer — SHAP latest documentation [Internet]. [cited 2026 Aug 23]. Available from: <https://shap-lrjball.readthedocs.io/en/latest/generated/shap.KernelExplainer.html>