Skin cancer detection using effective optical parameters and the classification and regression tree algorithm: A novel framework

Thanh Truc Nguyen[1, 2], Duc Minh Nguyen Huu[3], Thanh-Hai Le[4], Quoc-Hung Phan[5], Thi-Thu-Hien Pham[1, 2*]

[1]*School of Biomedical Engineering, International University, Vietnam National University Ho Chi Minh City, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University Ho Chi Minh City, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam*
[3]*Faculty of Traditional Medicine, University of Medicine and Pharmacy at Ho Chi Minh City, 217 Hong Bang, Ward 11, District 5, Ho Chi Minh City, Vietnam*
[4]*Department of Information Technology Specialization, FPT University, Lot E2a-7, Road D1 Hi-Tech Park, Long Thanh My Ward, District 9, Ho Chi Minh City, Vietnam*
[5]*Mechanical Engineering Department, National United University, 2, Lienda, Miaoli, Taiwan*

Received 20 July 2022; accepted 19 October 2022

Abstract:

Early detection of skin cancer matters because diagnosis, prognosis and treatment plan differ for each skin cancer type at their stages. Medical imaging taking the advantages of the non-invasive and non-ionizing polarized light is emerging as a tool for the development of screening and diagnotic tests. In this work, we proposed a novel framework to classify human melanoma and nonmelanoma skin cancer using the Classification and Regression Tree algorithm (CART). The samples were prepared from twenty-four non-melanoma skin cancer samples (consisting of twelve squamous cell carcinoma and twelve basal cell carcinoma samples); and three melanoma skin cancer samples. We calculated ten optical parameters from anisotropic biological tissues, namely the LB orientation angle (α), the LB phase retardance (β), the CB optical rotation angle (γ), the LD orientation angle (θd), the linear dichroism (D), the circular dichroism (R), the degrees of linear depolarization (e_1 and e_2), the degree of circular depolarization (e_3), and the depolarization index (Δ) using Stokes-Mueller matrix formalism. All effective optical parameters of biological tissue were then input into the CART classifier as predictors. The model yielded an accuracy of 92.6%, which is desirable for any robust and interpretable classification model. The results showed that for biological tissue samples, linear polarization properties dominate over circular ones due to the cellular microstructural composition of tissue, especially under anomalous growth as seen in skin cancer. This novel framework can potentially assist physicians in making timely and well-informed medical decisions.

Keywords: **classification and regression tree algorithm, human skin cancer, Stokes-Mueller matrix formalism.**

Classification numbers: **3.2, 3.6**

1. Introduction

Unfortunately, skin cancer has been rising at alarming rates, fortifying its position as the fifth-most common cancer in 2018 with more than 1.04 million cases of death recorded [1]. Characterized by uncontrolled growth and aggressive spread of skin tissues, skin cancer is categorized into malignant melanoma, squamous cell carcinoma, and basal cell carcinoma, which is listed by increasing frequency of diagnosis and decreasing level of severity. The latter two are collectively called non-melanoma skin cancer (NMSC) for obvious reasons. Both melanoma and NMSC are peculiar: the least common of the three - malignant melanoma - is the most lethal if not quickly diagnosed and treated. Many techniques have been developed for skin cancer detection and diagnosis such as dermatoscopy, reflectance confocal microscopy (RCM), nonlinear optical microscopy, and optical coherence tomography (OCT). Dermatoscopy, also known as epiluminescence microscopy, is a standard physical procedure where the area of the suspected lesion becomes magnified under a bright white light with an occasional application of oil or alcohol to improve the vision field [2]. However, some disadvantages of this imaging technique include a high dependence on the expertise and experience of the examiner as well as a high dependence on the appearance of classic dermoscopic features, which limits early diagnosis or any diagnosis of featureless melanoma [3]. RCM is an optical imaging technique that allows the skin to be analysed with nearly histological resolution at a cellular level [4-5]. Similar to dermoscopy images, real-time images obtained by RCM are horizontally oriented to the skin surface. RCM has been applied in medical settings for the diagnosis of melanoma and non-melanoma lesions and proven to increase diagnostic accuracy when used with dermoscopy [6]. Besides its application in skin oncology, RCM can be useful to delineate indications for inflammatory and infectious skin conditions. The main limitation of RCM is its relatively low skin penetration due to strong light scattering. A penetration depth of ~200 µm can be achieved, which is not sufficient to image in the dermis [6]. Also problematic is the fact that RCM sections are oriented perpendicular to conventional histological sections, making them difficult to interpret. OCT is a non-invasive, cross-sectional, real-time technique that allows conclusions to be drawn with regard to the presence of pathologies. Full-field OCT (FF-OCT) is one particular approach of OCT based on white-light interference microscopy that produces

*Corresponding author: Email: ptthien@hcmiu.edu.vn

en face tomographic images by the arithmetic combination of interferometric images acquired with an area camera and by illuminating the whole field of view with low-coherence light, making it suitable for high spatial resolution (~1.0 μm) imaging in both lateral and axial directions [7]. Line-field OCT (LF-OCT) uses a broadband spatially coherent light source and a line-scan camera in an interference microscope, providing a significant advantage over FF-OCT in terms of imaging penetration depth due to the confocal gate achieved by line illumination and detection [8-9]. S. Batz, et al. (2018) [10] showed that it is possible to distinguish between different nonmelanoma skin cancers by using OCT, but further prospective studies should be conducted to validate the sensitivity and specificity of the criteria. However, melanoma and pigmented elements proved to be a challenge for OCT as confirmed in [11]. Nonlinear optical microscopy is a high-resolution imaging modality based on nonlinear interactions of light with biological tissues [12]. Compared with OCT, nonlinear microscopy offers better spatial resolution, similar to that of RCM. Advances in developing multiphoton excitation microscopes with novel contrast mechanisms, such as second harmonic generation, further allow visualization of skin morphology and function [13]. Key limitations of nonlinear optical microscopy are the orientation of the images (en face sections, like RCM), the small field of view (350 μm × 350 μm with the DermaInspect device, Jenlab), and the relatively weak penetration in skin (~200 μm). Thus, a reliable and simple technique for skin cancer diagnosis applications is still needed.

Classification tree has long been used in biomedical engineering to determine a patient's prognosis after heart attack [14]. H. Henning, et al. (1979) [15] tried to identify patients who are at risk of death within 30 days among those that survived a heart attack for at least 24 hours after hospital admission using the CART classification algorithm. The problem of distinguishing early death from survivors with complete 19-variable from 215 patients produced a simple CART classification tree using only 3 variables. Ten-fold cross validation was employed and the model's performance was evaluated using an adjusted re-substitution of 0.21. L. Goldman, et al. (1982) [16] also used the CART classification algorithm to determine if patients with chest muscle pain would suffer from a heart attack. With an indicative electrocardiogram, characteristic elevation of pain enzyme level, and pain description from 482 patients, CART algorithm produced a much more complicated 13-split classification tree with pruning done based on sensitivity and specificity. The re-substitution overall misclassification cost was 0.07. The CART classification algorithm was also applied to classify immuno-suppression in patients with cancer using 21 continuous variables. The classification tree was pruned to minimize the re-substitution misclassification cost to 0.22 after cross-validation. This study also points out that newer, more expensive tests were not helpful as supplemental information and the variable thought to be a powerful diagnostic tool turned out to be one of the least important variables. These studies mostly deal with a large number of observations, an average number of variables, and only two class labels. The CART classification algorithm has been proven to outperform logistic classification and linear discrimination concerning sensitivity and specificity. CART is also used to detect outliers in a large number of observations and variables. Gait analysis done by D.H. Sutherland, et al. (1980) [17] used data with over 100 variables of 500 children to produce a classification tree with at least 16 terminal nodes. CART detected and helped correct two mistakes. This algorithm determined the relationship of variables to the development of mature gait. Dealing with a large number of variables is not uncommon, and principal component analysis is commonly used during dimensionality reduction before any further analysis. M.Z.F. Nasution, et al. (2018) [18] trained a classification tree using a C4.5 algorithm to detect patients with cervical cancer. The author used principal component analysis to reduce the number of variables from 36 to 12 relevant variables that explained 95% of the variance in the data set. With principal component analysis-based dimensionality reduction, the accuracy was 90.5% compared with an 86.05% rate for a tree without dimensionality reduction. High classification accuracy rates in frameworks with principal component analysis-based dimensionality reduction are reported in [19, 20]. Recently, the present group applied random forest algorithm to classify human skin cancer utilizing 16 Mueller matrix elements. The classifier obtained an average precision of 93% with the lowest score of 81% for basal cell carcinoma and the highest score of 100% for melanoma and squamous cell carcinoma [21]. F.V. Felix, et al. (2020) [22] proposed a machine learning-based method to detect tumours and healthy biological tissues with parameters that were extracted from optical diffuse reflectance spectroscopy. In this work, CART achieved the second highest accuracy of 93% in five applied machine learning models. Besides, M.S. Nogueira, et al. (2021) [23] used the diffuse reflectance spectra technique to extract optical tissue parameters for a training decision tree algorithm to recognize colorectal cancer. The tissue classification model achieved an average accuracy of 87% for both cancer detecting probes.

The present group previously proposed a decoupled analytical technique based on Stokes-Mueller matrix formalism for the successful extraction of all effective properties of turbid media and anisotropic optical samples [24, 25]. The proposed method was applied to extract linear birefringence (LB), linear dichroism (LD), circular birefringence (CB), circular dichroism (CD), linear depolarization (L-Dep), and circular depolarization (C-Dep) properties of various samples including human blood plasma, collagen, and calfskin [26], as well as squamous cell carcinoma and normal tissue in mice [27]. In this study, the classification of human skin cancer was performed using Stokes-Mueller matrix formalism and a CART classification algorithm. The practical feasibility of the proposed technique is examined by characterizing three types of skin cancer samples, namely, squamous cell carcinoma (SCC), basal cell carcinoma (BCC), and melanoma skin cancer (MSC).

2. Stokes-Muller matrix formalism

The decoupled analytical technique of T.T.H. Pham and Y.L. Lo has been previously described [24, 25]. This method can

extract ten effective parameters including the LB orientation angle (α), LB phase retardance (β), CB optical rotation angle (γ), LD orientation angle (θ_d), linear dichroism (D), circular dichroism (R), degrees of linear depolarization (e_1 and e_2), degree of circular depolarization (e_3), and depolarization index (Δ).

$$S_c = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = M_\Delta M_{lb} M_{cb} M_{ld} M_{cd} \hat{S}_c = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix} \begin{pmatrix} \hat{s}_0 \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \end{pmatrix}_c$$

where M_Δ, M_{lb}, M_{cb}, M_{ld}, and M_{cd} are the Mueller matrices for the depolarization, LB, CB, LD, and CD properties of the sample, respectively, and $\hat{s}_c$ is the input Stokes vector. In the methodology adopted in this study, the sample is radiated by four input linear polarization states, namely, $\hat{s}_{0^\circ} = [1 \;\; 1 \;\; 0 \;\; 0]^T$, $\hat{s}_{45^\circ} = [1 \;\; 0 \;\; 1 \;\; 0]^T$, $\hat{s}_{90^\circ} = [1 \;\; -1 \;\; 0 \;\; 0]^T$, and $\hat{s}_{135^\circ} = [1 \;\; 0 \;\; -1 \;\; 0]^T$) and two input circular polarization states, namely, right-handed $\hat{s}_{R-} = [1 \;\; 0 \;\; 0 \;\; 1]^T$ and left-handed $\hat{s}_{L-} = [1 \;\; 0 \;\; 0 \;\; -1]^T$.

Full details of the experimental procedure used to extract the various parameters are mentioned in Ref. [24, 25]. This methodology does not require the alignment of the principal birefringence and diattenuation axes. Although only four input polarization states (0°, 45°, 90°, and R-) are sufficient for obtaining all elements of the Mueller matrix, an extra two input polarization states (135° and L-) further improve the experimental results. Moreover, the ability of the analytical model to extract all optical parameters of interest over the measurement range was verified using analytical simulations and error analysis. Thus, the analytical model yielded accurate results even when the output Stokes parameters had errors in the range of ±0.005 or the samples had the minimum measurement of birefringence or dichroism [24, 25].

3. Experimental setup

The polarized light measurement system was set up horizontally to ensure laser source stability and safety during the experiments. This system includes a helium neon laser (wavelength of 632.8 nm, power < 5 mW), a quarter-wave plate, polarizers, and a Stokes polarimeter to characterize the linear birefringence, linear dichroism, circular 58 birefringence, circular dichroism, circular depolarization, and circular depolarization properties of a biological anisotropic sample. A frequency-stable He-Ne laser (HNLS008R, Thorlabs Co.) with a central wavelength of 633 nm served as source for input polarized light. A polarizer (GTH5M, Thorlabs Co.) and a quarter-wave plate (QWP0-63304-4-R10, CVI Co.) produced linear polarization light at 0, 45, 90 and 135 degrees; and two circular right-handed and left-handed polarizations. The optics were hooked up to a stage controller connected to a computer to automate the adjustment of polarizer angle. A neutral density filter (NDC-100-2, ONSET Co.) was used to ensure that each of the input polarization lights had the equal intensities. The output Stokes parameters were computed from intensity measurements obtained using a Stokes polarimeter (PAX5710, Thorlabs Co.) at a sampling rate of 33.33 samples per second. A minimum of 1024 data points were obtained for each sample and used to calculate the value of each effective parameter. The system was calibrated against a Poincare sphere after every run by returning the value of the angle to zero before setting the angle to the next value. The arrangement of the measurement system is shown in Fig. 1.

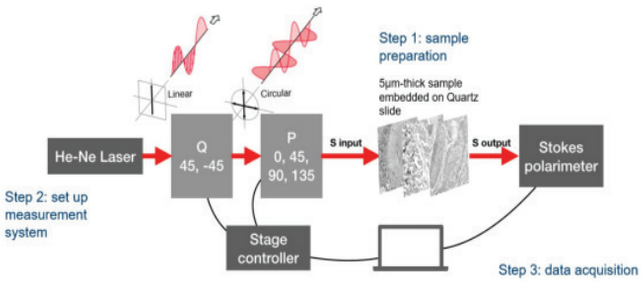

Fig. 1. The illustration of measurement system.

3.1. Sample preparation

Twenty-four non-melanoma skin cancer samples (consisting of twelve SCC and twelve BCC samples) and three MSC samples were acquired from Khanh Hoa General Hospital (Nha Trang city, Vietnam). The sample labels, which were affixed by the hospital, included information on the type of cancer, the stage of the disease, and the date of packaging. All samples were embedded and fixed in formalin and paraffin. The samples were sectioned with a thickness of 5 μm by a microtome and placed on quartz slides. Based on Ref. [28], before cutting samples, the thicknesses of the samples were recorded. After the calibration, the 5-μm thick samples were chosen to ensure both the sensitivity of the received polarized light at the Stokes polarimeter and the ability to characterize the polarization properties of samples. Furthermore, cancerous tissue samples were stained with H&E and observed under an optical microscope (OM) for histopathological analysis.

3.2. Dimension reduction

Principal component analysis reduces data dimensionality by extracting principal components that reflect relevant features in the data [29-30]. The benefit is that a significant proportion of the variance in the data can be explained by a reduced number of orthogonal components compared to the total number of raw input variables. Principal component analysis was performed by singular value decomposition. The values of the variable were standardized before performing principal component analysis by subtracting the values of each variable by the variable's mean and scaling them with the variable's standard deviation. We included a line in the code to determine if any of the categorical variables were excluded out of the principal component analysis because only numeric variables can be broken down to principal components. The dimensions that explain 95% of the variance

in the data were enough to summarize the maximum amount of information contained in the original set of optical parameters. Sixty-one PCA results were tested for sampling adequacy using a Kaiser-Meyer-Olkin test followed by Bartlett's test of sphericity to test whether the covariance matrix is significantly different from identity [31].

3.3. Training

The predictors used to train the CART classification algorithm are the dimensions selected by principal components analysis. The predictor values are the principal components scores of the dimensions explaining 95% of the variance existing in the data set. The CART algorithm requires the split criterion to be based on the Gini diversity index. Because there is no instance of missing data, surrogate splits were not designated to handle missing data. The prune criterion is let to error. The split predictor was selected to maximize the Gini diversity index over all possible splits of all predictors following the construction of the CART classifier algorithm. The original predictors were passed through a principal component analysis and the new predictor was the principal component scores, hence, they do not reflect the nature of the optical parameter predictors. The maximum number of splits were run from 1 to 10 to produce trees ranging from coarse to fine to see the rate of change in performance as the trees gain more leaves.

3.4. Cross-validation

Ten-fold cross-validations were performed on the classification model and initialized the prediction to the proper size of each fold. The data set was randomly partitioned into 10 folds. To ensure reproducibility, a random number generator was used. For each fold, the process of cross validation mimics the steps described above: predictors and responses of each fold are defined; principal component analysis was conducted on the numeric predictors matrix; the dimensions that cumulatively explain 95% of variance of the data set are kept; a classifier is trained with hyper-parameters described in the training 62 section; and, finally, the prediction is calculated together with the fold score. After going through 10 folds, the correct predictions were identified by comparing fold predictions to the correct responses and validation accuracy was calculated by taking the mean of the correct predictions of all folds. The cross-validation scheme is presented in Fig. 2.

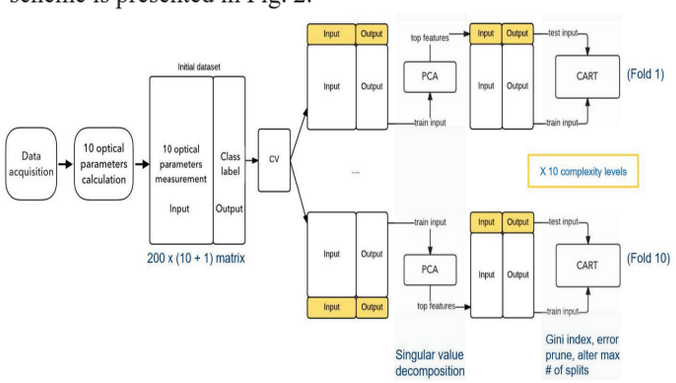

Fig. 2. The cross-validation scheme.

4. Results and discussion

Table 1 shows the values of ten effective optical parameters of normal and cancerous skin samples. The results showed that α presented a degree of differential power across the four sample types with normal skin tissue ranked highest at 15.14 degrees. The same trend was observed for β, γ, and D with normal skin ranking highest at 1.47 degrees, 0.24 degrees, and 0.12, respectively. The opposite trend with normal skin tissue ranked lowest was observed in θ_d of 97.85 degrees. It can be asserted with a high level of confidence that the ten effective optical parameters derived with proposed technique have the ability to differentiate normal human skin tissue from cancerous human skin tissue.

Table 1. Results of ten effective parameters.

		α	β	γ	θ_d	D	R	e_1	e_2	e_3	Δ
Normal skin	Mean	15.14	1.43	0.24	97.85	0.12	-0.04	$3.7x10^{-6}$	$3.9x10^{-6}$	$1.1x10^{-5}$	0.999
	SD*	2.02	0.13	0.03	4.90	0.01	0.008	$2.4x10^{-7}$	$2.3x10^{-6}$	$7.0x10^{-6}$	$4.4x10^{-6}$
Squamous cell carcinoma	Mean	3.05	0.35	0.05	124.49	0.08	-0.022	$9.6x10^{-6}$	$9.5x10^{-6}$	$3.4x10^{-5}$	0.999
	SD*	0.85	0.03	0.007	4.81	0.01	0.002	$2.9x10^{-7}$	$2.6x10^{-7}$	$1.2x10^{-6}$	$7.2x10^{-6}$
Basal cell carcinoma	Mean	7.03	0.55	0.054	125.31	0.05	-0.023	$2.1x10^{-6}$	$3.4x10^{-6}$	$1.6x10^{-5}$	0.999
	SD*	0.94	0.03	0.006	2.58	0.005	0.002	$6.2x10^{-7}$	$8.4x10^{-7}$	$3.2x10^{-6}$	$1.6x10^{-6}$
Melanoma	Mean	5.47	0.44	0.09	126.58	0.07	0.008	$1.7x10^{-6}$	$1.6x10^{-6}$	$1.7x10^{-5}$	0.999
	SD*	0.37	0.02	0.01	3.78	0.01	0.001	$4.6x10^{-7}$	$4.1x10^{-7}$	$3.6x10^{-6}$	$2.1x10^{-6}$

*SD = Standard deviation.

4.1. Extracting the principal components

Principal component analysis was used to avoid overfitting by reducing the number of effective optical parameters used as predictors for the CART classification algorithm as illustrated as Fig. 3. As shown, two principal components were yielded from the dataset of 10 different effective optical parameters, which explained 95% of the dataset variance. The first principal component explained the largest proportion, 74.4% of the variance, and was composed of linear optical parameters θ_d and Δ with 76.22 and 27.12%, respectively. The second principal component explained the second largest proportion of 20.6% of the variance and was composed of linear optical parameters with α and β positively contributing 64.23 and 16.67%, respectively.

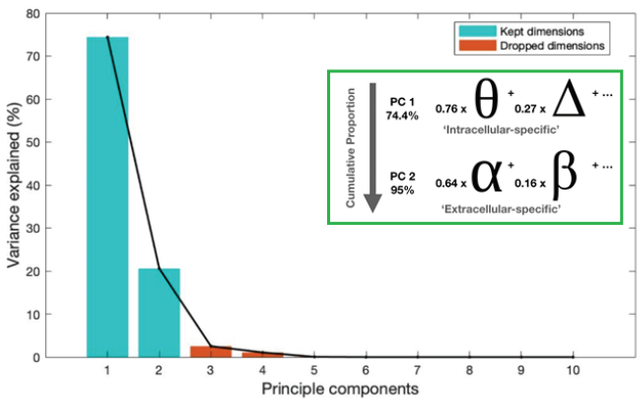

Fig. 3. Breaking down the principal components used as predictors for the CART classifier.

Figure 4 shows the data distributions of the optical parameters α, β, θ_d, and Δ of four skin cancer samples including normal, melanoma, BCC, and SCC.

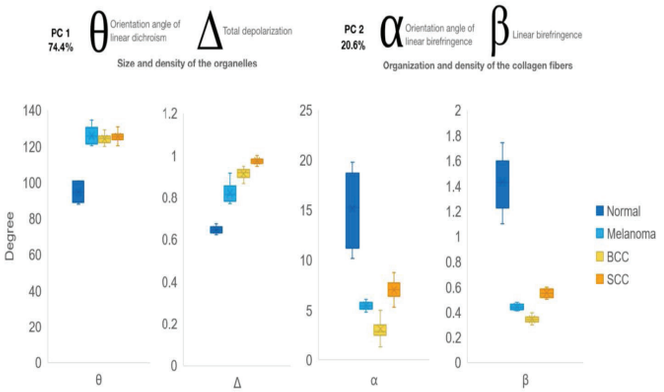

Fig. 4. Breaking down the metrics that matter.

4.2. Evaluation of the CART classification algorithm

The maximum number of splits were run from 1 to 10 to examine the performance of the classification algorithm at increasing levels of complexity, which is determined by the number of nodes and branches. Fig. 5 shows the illustration of the CART classification tree with the maximum number of splits. As shown, the classification ended up with 5 levels of tree complexity at 5 different maximum number of splits, specifically, 1 (Fig. 5A); 2 (Fig. 5B); 3 (Fig. 5C); 4, 5, 6 (Fig. 5D); and 7, 8, 9, 10 (Fig. 5E). The CART classification algorithm returned the same tree, meaning a tree with the same number of splits and nodes, same split predictors, same split points for configurations with a maximum of 4, 5, and 6 splits (Fig. 5C), and applied with configurations with a maximum 7, 8, 9 and 10 splits (Fig. 5D). This is because the Gini index reached a threshold of maximum knowledge gain given the number of splits. The 5 classification trees were evaluated by their training accuracy, testing accuracy, error rate, recall or sensitivity, precision, specificity, and F1 score shown in Table 2. The CART classification tree with a maximum number of splits = 4 performed with the best F1 score of 98.59% describing the overall performance of the models by finding the common ground between recall and precision. F1 is a good indicator when a dataset is unbalanced, such as the dataset with which we were dealing. It is noted that, for the CART classification tree with a maximum number of splits = 1, precision was NaN because the classification tree did not include enough classes to determine precision. Therefore, because F1 reflects precision and recall, it returned as NaN.

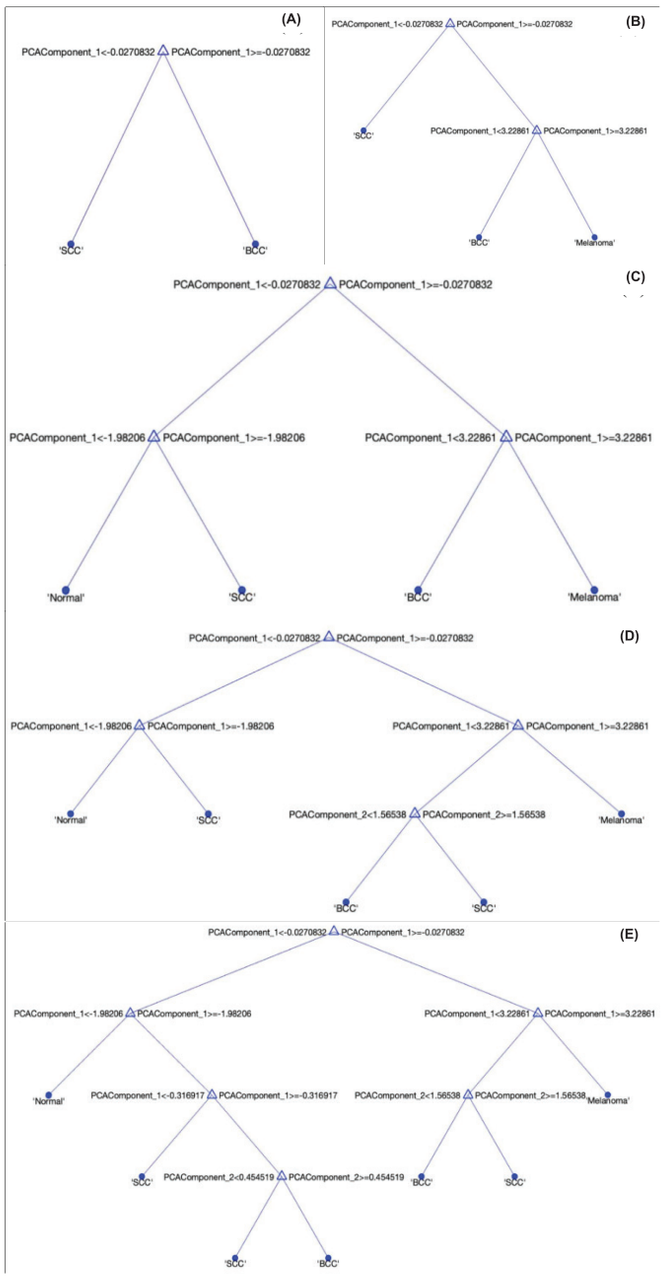

Fig. 5. CART classification tree with maximum number of splits = (A) 1, (B) 2, (C) 3, (D) 4, 5, 6 and (E) 7, 8, 9, 10.

Table 2. Performance evaluation of different CART trees.

Maximum of number of splits	1	2	3	4	7
Training accuracy	81.4%	87.7%	92.9%	93.1%	92.7%
Testing accuracy	78.4%	85.8%	91.3%	92.6%	92.1%
Error rate	20%	16.2%	8.7%	7.8%	8.7%
Recall (Sensitivity)	0%	48.6%	97.1%	100%	97.1%
Precision	NaN	94.4%	97.1%	97.2%	97.1%
Specificity	100%	98.8%	98.8%	99.8%	98.8%
F1	NaN	64.15%	97.1%	98.59%	97.1%

Figure 6 shows the performance evaluation of different CART trees, highlighting the superiority of MaxSplits4 with a maximum number of splits = 4. The capability of this framework as a trustful tool for decision making is shown by a high accuracy of 94% at high speed and low computing cost. With a decision tree-based CART classification algorithm, the visualization of the decision making process regarding which observation belongs to which cancer type is straight forward in logic flow, reflecting the resemblance between the CART classification algorithm and a basic human decision making process. In other words, the decision of which predictor should be used for a split and where to split depends on the extent of the distinguishing impact of the split selection measured by the impurity of the children nodes. Using this decision tree produced by CART, the effective optical parameters that play the most important role in representing the physical properties of the biological anisotropic tissue samples can be determined. It is important to note that with different numbers of features selected as predictors and different configurations specified for CART (most significantly the number of maximum nodes), the most important predictors yielded can vary, which is not necessarily the predictors of the first splits [32].

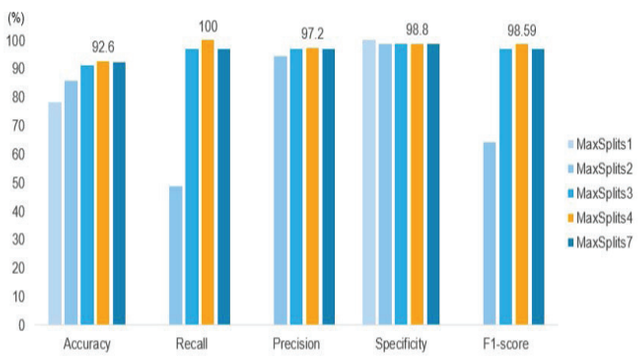

Fig. 6. Performance evaluation of different CART trees, highlighting the superiority of MaxSplits4 (orange) with a maximum number of splits = 4.

When performing the experiments, although the visual image of the Poincare sphere is quite clear, manually adjusting the laser light path to be perpendicular to the optical lens and filter is not guaranteed to be exact. This leads to an average skewness in the ellipticity of 1 to 2 degrees away from the exact 0, 45, 90, and135-degree targets. Despite each optical parameter being derived from a mathematical formula to obtain an insignificant difference from the accurate target angle of ellipticity and degree of polarization, sequences of distorted results can originate from the system to create an cumulative error when large-scale data is collected. This problem makes it difficult to repeat the experiment for a robust sampling plan. Therefore, constant and accurate calibration and noise factor identification are essential factors to develop a database that delivers proper diagnostics on skin samples in a large-scale clinical investigation in the future. A quick comparison to the ideal state of the linear polarizer will help users reassess the experiment and recalibrate the system. Basically, this study indicates a comprehensive collection of effective parameters of normal skin, which can be used as a reference for further research. The method of using a Muller matrix in the medical/healthcare system is an approach for detecting cancer cells by optical parameters obtained from the measurements of normal and carcinoma human skin tissue. More structural organization and characteristics of human skin tissue can be obtained under the perspective of crude positions of the organelles and the extracellular matrix while reserving the high interpretability of the CART classifier and enhancing its accuracy, which helps differentiate normal, squamous cell carcinoma, basal cell carcinoma, and melanoma.

5. Conclusions

The Stokes-Mueller method allows the extraction of ten structure-dependent effective optical parameters. Using principal component analysis and a highly interpretable CART classification algorithm framework proposed in this work, melanoma, non-melanoma, and healthy skin tissue were able to be classified with an accuracy of 92.6%. The results also showed that linear optical properties dominated the biological anisotropic samples. The investigation into human skin cancer revealed that intracellular-specific properties, driven mainly by linear dichroism, and extracellular specific properties, driven by linear birefringence, are strong indicators of anisotropy and anomalies found in cancer tissues. This framework can potentially assist physicians in making timely and well-informed medical decisions that save lives.

CRediT author statement

Thanh Truc Nguyen: Software, Investigation, Data curation, Writing draft; Duc Minh Nguyen Huu: Visualization, Investigation, Review & Editing; Thanh-Hai Le: Software, Visualization, Investigation, Data curation; Quoc-Hung Phan: Conceptualization, Methodology, Validation; Thi-Thu-Hien Pham: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing - Review & Editing.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided to this study by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 103.03-2019.381.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] J.S. Lin, et al. (2011), "Behavioral counseling to prevent skin cancer: A systematic review for the U.S. preventive services task force", *Annals of Internal Medicine*, **154(3)**, pp.190-201.

[2] P.A. Ascierto, et al. (2000), "Sensitivity and specificity of epiluminescence microscopy: Evaluation on a sample of 2731 excised cutaneous pigmented lesions", *Br. J. Dermatol.*, **142(5)**, pp.893-898.

[3] G. Argenziano, et al. (2008), "Dermoscopic monitoring of melanocytic skin lesions: Clinical outcome and patient compliance vary according to follow-up protocols", *Br. J. Dermatol.*, **159(2)**, pp.331-336.

[4] I. Alarcon, et al. (2014), "Impact of in vivo reflectance confocal microscopy on the number needed to treat melanoma in doubtful lesions", *Br. J. Dermatol.*, **170(4)**, pp.802-808.

[5] J. Champin, et al. (2014), "*In vivo* reflectance confocal microscopy to optimize the spaghetti technique for defining surgical margins of lentigo maligna", *Dermatologic Surg.*, **40(3)**, pp.247-256.

[6] P. Guitera, et al. (2010), "The impact of in vivo reflectance confocal microscopy on the diagnostic accuracy of lentigo maligna and equivocal pigmented and nonpigmented macules of the face", *J. Invest. Dermatol.*, **130(8)**, pp.2080-2091.

[7] J. Olsen, et al. (2018), "Advances in optical coherence tomography for dermatology-a review", *J. Biomed. Opt.*, **23(4)**, pp.1-10.

[8] S.H. Tsang, T. Sharma (2018), "Optical coherence tomography", *Advances in Experimental Medicine and Biology*, **1085(5035)**, pp.11-13.

[9] R.F. Spaide, et al. (2018), "Optical coherence tomography angiography", *Prog. Retin. Eye Res.*, **64**, pp.1-55.

[10] S. Batz, et al. (2018), "Differentiation of different nonmelanoma skin cancer types using OCT", *Skin Pharmacol. Physiol.*, **31(5)**, pp.238-245.

[11] L.F.di Ruffano, et al. (2018), "Optical coherence tomography for diagnosing skin cancer in adults", *Cochrane Database Syst. Rev.*, **12**, DOI: 10.1002/14651858.CD013189.

[12] W.R. Zipfel, et al. (2003), "Nonlinear magic: Multiphoton microscopy in the biosciences", *Nat. Biotechnol.*, **21(11)**, pp.1369-1377.

[13] K. Koenig, I. Riemann (2003), "High-resolution multiphoton tomography of human skin with subcellular spatial resolution and picosecond time resolution", *J. Biomed. Opt.*, **8(3)**, pp.432-439.

[14] E.G.R. Olshen, et al. (1983), "Risk prediction after myocardial infarction: Comparison of three multivariate methodologies", *Cardiology*, **70(2)**, pp.73-84.

[15] H. Henning, et al. (1979), "Prognosis after acute myocardial infarction: A multivariate analysis of mortality and survival", *Circulation*, **59(6)**, pp.1124-1136.

[16] L. Goldman, et al. (1982), "A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain", *N. Engl. J. Med.*, **307(10)**, pp.588-596.

[17] D.H. Sutherland, et al. (1980), "The development of mature gait", *J. Bone Joint Surg. Am.*, **62(3)**, pp.336-353.

[18] M.Z.F. Nasution, et al. (2018), "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification", *J. Phys.: Conf. Ser.*, **978**, DOI: 10.1088/1742-6596/978/1/012058.

[19] Y. Zhang, et al. (2014), "Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree", *Prog. Electromagn. Res.*, **144**, pp.171-184.

[20] E. Gokgoz, A. Subasi (2015), "Comparison of decision tree algorithms for EMG signal classification using DWT", *Biomed. Signal Process. Control*, **18**, pp.138-144.

[21] N.T. Luu, et al. (2021), "Characterization of Mueller matrix elements for classifying human skin cancer utilizing random forest algorithm", *J. Biomed. Optics*, **26(7)**, DOI: 10.1117/1.JBO.26.7.075001.

[22] F.V. Felix, et al. (2020), "Application of classification algorithms to diffuse reflectance spectroscopy measurements for ex vivo characterization of biological tissues", *Entropy (Basel)*, **22(7)**, DOI: 10.3390/e22070736.

[23] M.S. Nogueira, et al. (2021), "Tissue biomolecular and microstructure profiles in optical colorectal cancer delineation", *J. Phys. D: Appl. Phys.*, **54(45)**, DOI: 10.1088/1361-6463/ac1137.

[24] T.-T.-H. Pham and Y.-L. Lo (2012a), "Extraction of effective parameters of turbid media utilizing the Mueller matrix approach: study of glucose sensing," *J. Biomed. Opt.*, **17(9)**, DOI: 10.1088/1361-6463/ac1137.

[25] T.T.H. Pham, Y.L. Lo (2012b), "Extraction of effective parameters of anisotropic optical materials using a decoupled analytical method", *J. Biomed. Opt.*, **17(2)**, DOI: 10.1117/1.JBO.17.2.025006.

[26] H.T.T. Pham, et al. (2018), "Optical parameters of human blood plasma, collagen, and calfskin based on the Stokes-Mueller technique", *Applied Optics*, **57(16)**, pp.4353-4359.

[27] D.L. Le, et al. (2018), "Characterization of healthy and nonmelanoma-induced mouse utilizing the Stokes-Mueller decomposition", *J. Biomed. Opt.*, **23(12)**, pp.1-8.

[28] H.R. Lee, et al. (2019), "Digital histology with Mueller microscopy: How to mitigate an impact of tissue cut thickness fluctuations", *J. Biomed. Opt.*, **24(7)**, DOI: 10.1117/1.JBO.24.7.076004.

[29] I.T. Jolliffe (2002), "Graphical representation of data using principal components", *Principal Component Analysis, Springer Series in Statistics*, Springer, pp.78-110.

[30] H. Abdi (2004), "Linear Algebra for neural networks", *International Encyclopedia of the Social & Behavioral Sciences*, pp.1-9.

[31] C.D. Dziuban, E.C. Shirkey (1974), "When is a correlation matrix appropriate for factor analysis? Some decision rules", *Psychol. Bull.*, **81(6)**, pp.358-361.

[32] L. Breiman, et al. (2017), "Classification and regression trees", *Routledge*, 368pp.