

ENHANCING MULTI-CLASS PREDICTION OF SKIN LESIONS WITH FEATURE IMPORTANCE ASSESSMENT

AGNE PAULAUŠKAITE-TARASEVICIENE ^{a,*}, KRISTINA SUTIENE ^b, NOJUS DIMSA ^c,
SKAIDRA VALIUKEVICIENE ^{d,e}

^aArtificial Intelligence Centre
Kaunas University of Technology
K. Barsausko g. 59, 51423 Kaunas, Lithuania
e-mail: agne.paulauskaite-taraseviciene@ktu.lt

^bDepartment of Mathematical Modeling
Kaunas University of Technology
Studentu g. 50, 51368 Kaunas, Lithuania
e-mail: kristina.sutiene@ktu.lt

^cFaculty of Informatics
Kaunas University of Technology
Studentu g. 50, 51368 Kaunas, Lithuania
e-mail: nojus.dimsa@ktu.edu

^dDepartment of Skin and Venereal Diseases
Lithuanian University of Health Sciences
A. Mickevičiaus g. 9, 44307 Kaunas, Lithuania
e-mail: skaidra.valiukeviciene@kaunoklinikos.lt

^eDepartment of Skin and Venereal Diseases
Hospital of Lithuanian University of Health Sciences 'Kauno klinikos'
Eiveniu g. 2, 50161 Kaunas, Lithuania

Numerous image processing techniques have been developed for the identification of various types of skin lesions. In real-world scenarios, the specific lesion type is often unknown in advance, leading to a multi-class prediction challenge. The available evidence underscores the importance of employing a comprehensive array of diverse features and subsequently identifying the most important ones as a crucial step in visual diagnostics. For this purpose, we addressed both binary and five-class classification tasks using a small dataset, with skin lesions prevalent in Lithuania. The model was trained using a rich set of 662 features, encompassing both conventional image features and graph-based ones, which were obtained from the superpixel graph generated using Delaunay triangulation. We explored the influence of feature importance determined by SHAP values, resulting in a weighted F1-score of 92.48% for the two-class classification and 71.21% for the five-class prediction.

Keywords: skin lesion, feature extraction, graph theory, multi-class prediction, SHAP values.

1. Introduction

Skin cancer ranks among the most lethal forms of cancer globally, contributing significantly to mortality

rates worldwide. Early detection plays a pivotal role in mitigating fatalities attributed to skin cancer. However, the conventional diagnostic method (visual inspection) is often not effective enough. To address this challenge, artificial intelligence approaches have emerged

*Corresponding author

as promising aids for dermatologists in achieving timely and precise diagnoses of skin cancers. Progress in skin cancer detection continues to be made systematically, with various technologies and methodologies being developed to improve accuracy and accessibility. Different machine learning techniques including decision trees (Vikas Reddy and Rama Parvathy, 2022), random forests (Damian *et al.*, 2022), support vector machines (Murugan *et al.*, 2019) or K-nearest neighbors (Abbes *et al.*, 2021) are being investigated, with the primary focus on essential feature extraction. However, it is not surprising that deep learning-based imaging techniques have become increasingly prevalent in this field (Tembhurne *et al.*, 2023; Dildar *et al.*, 2021) like in all other medical imaging tasks. Recent studies have shown that deep learning algorithms implemented into contemporary smartphone cameras enable self-examinations of skin lesions achieving a high sensitivity and specificity in the classification of melanomas and melanocytic nevi (Liutkus *et al.*, 2023).

The computer-vision powered diagnosis of skin cancer typically involves five steps: image acquisition, preprocessing, segmentation, feature extraction, and classification. Among these, segmentation (Araújo *et al.*, 2021; Ashraf *et al.*, 2022; Oukil *et al.*, 2021; Surówka and Ogorzałek, 2022) and classification (Ali *et al.*, 2022; Aladhadh *et al.*, 2022) tasks have attracted close attention. However, achieving precise diagnosis with image deep learning algorithms is challenging and necessitates consideration of numerous factors. For instance, artifacts like hairs, dark corners, water bubbles, marker marks, ink marks, and ruler marks can result in misclassification and inaccurate segmentation of skin lesions. These tasks are therefore very sensitive to the size and quality of the dataset.

Although there are over 20 open-access datasets available, the lack of transparency in reporting metadata for clinically essential characteristics constrains the clinical utility of these images, particularly when they are reused across datasets (Wen *et al.*, 2022). Furthermore, machine learning algorithms utilized for medical image classification are recognized to perform inadequately on images collected from populations independent of those used for training (Navarrete-Dechent *et al.*, 2018). Nevertheless, the most popular open datasets such as ISBI 2016 (Gutman *et al.*, 2016) and ISBI 2017 (Codella *et al.*, 2017), PH2 (Mendonca *et al.*, 2013), ISIC (2016-2020) challenge datasets (ISIM-ISIC, 2020), BCN20000 (Combalia *et al.*, 2019), and HAM10000 (Tschandl *et al.*, 2018) provided in DICOM or JPEG formats, offer computer vision researchers invaluable resources for developing and evaluating algorithms focused on skin cancer detection, classification, and diagnosis, particularly emphasizing melanoma images. Among other skin cancer types, including basal cell

carcinoma (BCC), squamous cell carcinoma (SCC), and Merkel cell carcinoma (MCC), malignant melanoma is extensively researched due to its potentially aggressive nature and higher risk of metastasis compared with other types of skin cancer. Many studies have achieved high accuracy ($> 90\%$ ACC) in two-class detection tasks involving benign and malignant lesions (melanoma) (Hurtado and Reales, 2021). However, studies on imbalanced small datasets for multi-class skin lesion classification typically do not surpass 86% accuracy (Alwakid *et al.*, 2022; Rashid *et al.*, 2019; Abdelhalim *et al.*, 2021).

The accuracy of diagnoses achieved by each of AI-based approaches still lacks stability in results (as it heavily depends on the data), with only dermoscopy being commonly used by all dermatologists. Additionally, other methods such as confocal laser scanning microscopy, optical coherence tomography, 3D topography or multispectral imaging can serve as optical techniques for skin examination. Multispectral imaging entails capturing skin images across multiple wavelengths of light, unveiling diverse characteristics of skin lesions. Its potential lies in enhancing the accuracy of skin cancer diagnosis by furnishing supplementary information about the lesions (Ilişanu *et al.*, 2023; Rey-Barroso *et al.*, 2018). While there is a significant potential for conducting important studies on this type of images, there is currently no representative dataset available, resulting in relatively limited research being conducted.

In our study, we conducted a skin lesion detection, addressing both binary and five-class prediction challenges. All experiments were carried out using a dataset provided by the Lithuanian University of Health Sciences, focusing on skin lesions, namely naevus, seborrheic keratosis, melanoma, dermatofibroma, and lentigo malignant, most prevalent in the region of Lithuania. Notably, various types of skin lesions exist, and their prevalence depends on geographic location, sun exposure habits, skin type, immunosuppression, etc. However, a dataset used in the study was severely imbalanced, making it inherently challenging to discriminate between the various types of skin lesions. Among many possible ways to conduct the research (Zafar *et al.*, 2023), we focus on the feature extraction and their importance measuring for the skin lesion detection by formulating a multi-class prediction model. More specifically, the study aimed to evaluate the effectiveness of graph-based features derived from a superpixel graph generated using Delaunay triangulation, which were combined together with the conventional image features. A multi-class random forest (RF) was built to determine the feature importance in terms of both Mean Decrease Impurity (MDI) and SHapley Additive exPlanations (SHAP) values, while the localisation of skin lesion was conducted using YOLOv8 (Jocher *et al.*, 2023).

This choice was made due to a relatively small dataset used in the study and promising results of ensemble methods, when graph-based features were used (Annaby *et al.*, 2021; Oliveira *et al.*, 2017). Delaunay triangulation was employed due to its functionality to focus specifically on the cancerous lesion while neglecting the surrounding skin, as highlighted in the work by Sunarya *et al.* (2023). To sum up the experimental results, in total 662 features were extracted, among which 125 features represent the vertex domain, 24 features represent the spectral domain, and 513 are image conventional features. Based on MDI and SHAP threshold values, the proposed solution achieved a weighted F1 score of 92.48% for the two-class problem and 71.21% for the five-class problem, which is in line with many other studies published in this domain.

2. Methodology

2.1. Image-set description. The study utilizes images extracted using the SIAscope apparatus provided by Lithuanian Health Science University. The SIAscope technology emits harmless radiation into the skin within wavelengths of 400 to 950 nm, measuring the reflected light at each wavelength. This process exploits the distinct optical properties of different skin components, which absorb and reflect light differently, favoring certain wavelengths. From these spectral measurements, SIAscope derives information about the location, quantity, and distribution of melanin, collagen, and hemoglobin (blood vessels) within the skin layers (Emery *et al.*, 2010). The dataset comprises 43 samples classified as class 0 (naevus), 36 samples as Class 1 (seborrheic keratosis), 219 samples as Class 2 (melanoma), 30 samples as Class 3 (dermatofibroma), and 144 samples as Class 4 (lentigo malignant), totaling 472 specimens (see Fig. 1). All images we received were complete with no obvious artefacts inside. Each specimen comprises five images: a $1544 \times 1544 \times 3$ RGB melanin, a $708 \times 708 \times 1$ grayscale melanin, a $708 \times 708 \times 3$ hemoglobin, a $708 \times 708 \times 3$ collagen, and a $708 \times 708 \times 3$ derma melanin image.

This research focuses on exploring RGB images based on pre-test experiments, which have indicated that RGB images yield slightly better classification results compared with other types. More specifically, YOLOv7 (Wang *et al.*, 2022) and YOLOv8 (Jocher *et al.*, 2023) models were applied to localize the skin lesion as the region of interest (see Table 1). Results in Table 1 suggests that the best detection of localization of skin lesion is observed for RGB images. Therefore, the multi-class skin lesion classification model was employed only for RGB images.

2.2. Skin lesion classification model. The general idea of multi-class skin lesion detection model is depicted

Table 1. Mean average precision (mAP) for object detection results using YOLOv7 and YOLOv8.

Image type	YOLOv7		YOLOv8	
	mAP50	mAP50-95	mAP50	mAP50-95
RGB melanin	0.7825	0.5705	0.8891	0.6012
Grayscale melanin	0.6214	0.3812	0.7740	0.5269
Hemoglobin	0.4984	0.3274	0.8231	0.5081
Collagen	0.4003	0.2239	0.7114	0.4225
Derma melanin	0.3357	0.3011	0.4746	0.4046
Averaged image	0.6769	0.4637	0.8093	0.5851

in Fig. 2. For every input dermoscopic image (see Section 2.1), the corresponding binary mask was utilized to eliminate the background. This preprocessing step ensures that the focus remains solely on the region of interest, which is the skin lesion itself. Subsequently, all images were resized to dimension of 750×750 . Following this, the superpixel graph representation of an image was constructed (see Section 2.2.1). The conventional features from pixel information and graph-based features were extracted (see Section 2.2.2). Then, feature selection was employed, which relied on both MDI- and SHAP-based threshold values (see Sections 2.2.3 and 2.2.4). Those features were fed into the random forest multi-class classifier to predict the skin lesion type. The settings used in the experiments are as follows: number of estimators = 100, min sample split = 2, min sample leaf = 1, max features = \sqrt{d} , bootstrap = true, split quality measured by the Gini index.

Due to the significantly unbalanced dataset, we conducted the experiment by adjusting the proportions of training and testing samples, initially starting from 90% for training and 10% for testing, and then changing them every five percent to 60% and 40%, respectively. We observed that the best results were obtained with sample proportions of 70% and 30%, 65% and 35%, and 60% and 40% for training and testing. Given the highest accuracy obtained, the experimental results were provided in the paper using the 60% and 40% proportion.

The result of the classification model is the predicted type of skin lesion. For a binary classification model, Melanoma and lentigo malignant were attributed to Class 1 as being malignant, while naevus, seborrheic keratosis, and dermatofibroma were assigned to Class 0 as being benign. In the multi-class prediction model,

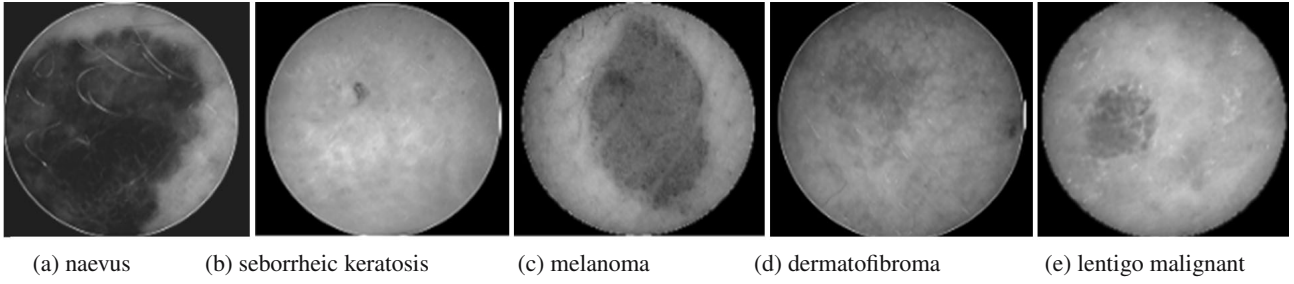


Fig. 1. Sample images for all five classes.

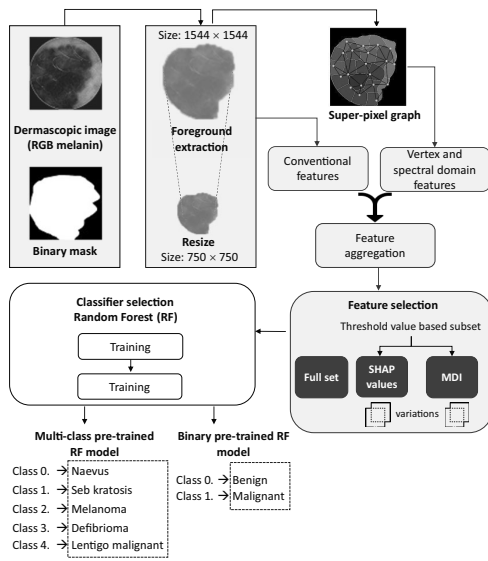


Fig. 2. Multi-class skin lesion classification model.

there are five distinct classes corresponding to each type of lesion.

2.2.1. Superpixel graph representation of skin lesion images. In the study, the superpixels were created following the idea proposed by Annaby *et al.* (2021). The pseudo-algorithm, which includes the steps of clustering and triangulation to generate a superpixel graph is represented as Algorithm 1.

More specifically, first, to reduce the complexity of image processing, individual pixels were clustered into groups that are called superpixels. For this purpose, the simple linear iterative clustering (SLIC) algorithm could be employed (Achanta *et al.*, 2010). Second, the technique proposed by Sharma *et al.* (2004) is used to decrease the number of superpixels to some specified value P , as the SLIC algorithm can result in varying numbers of superpixels for different images. In contrast to Annaby *et al.* (2021), Delaunay triangulation was employed in the generation of the graph. It enables the model to concentrate specifically on the cancerous

Algorithm 1. Superpixel graph generation.

Step 1. Load the image I and resize to 750×750 px

Step 2. Load the mask M in grayscale and resize to 750×750 px

Step 3. $S \leftarrow \text{SLIC}(I, M, \sigma, k)$ on the image with the mask to get segments, where k defines the number of clusters and σ denotes a width of Gaussian smoothing kernel

Step 4. Initialize an array A to store the graph vertex pixels of each segment $s \in S$

For each unique segment $s \in S$:

Select valid pixels $P_s = s \cap M$, where $P_s \neq \emptyset$

Determine a border B_s of P_s

Calculate the distance $D_s = \|P_s - B_s\|_2$

Select the pixels P_s^d , where $D_s \geq n$ units, where $n = 20$

If $P_s^d \neq \emptyset$ then $A = \text{append}(A_s)$, where A_s is a randomly selected pixel from P_s^d

Step 5. $G \leftarrow \text{DelaunayTriangulation}(A)$

lesion while disregarding the surrounding skin (Sunarya *et al.*, 2023). When generating graphs using Delaunay triangulation, each point, representing the superpixel, corresponds to a vertex in the graph, and each edge in the triangulation corresponds to an edge in the graph. In the end, we get a complete superpixel graph where every pair of distinct vertices is connected by a unique edge. The examples of generated superpixel graphs for all types of skin lesions are displayed in Fig. 3.

2.2.2. Feature extraction. Two main groups of features were extracted and later fed into the prediction model.

The first group consists of structural graph features that include time-domain features and frequency-domain features. Time-domain features define both local and global graph aspects that, in general, assess the graph complexity. In line with work (Annaby *et al.*, 2021), 5 global features and 6 local features were computed, which results in a total of $5 + 6 \times P$ feature values, where $P = 20$. Frequency-domain features were obtained

using the graph Fourier transform, which is based on the eigendecomposition of the graph Laplacian

$$L = U\Lambda U^T, \quad (1)$$

where $L = D - A$, A is the graph adjacency matrix, D is the diagonal degree matrix, Λ is the diagonal matrix of eigenvalues, and the columns of U are the eigenvectors. More specifically, the eigenvectors imply the orthonormal Fourier basis for signals, while the corresponding eigenvalues define graph frequencies. Together with additional characteristics from the graph Fourier transform such as energy, power, entropy, and amplitude, a feature vector of 24 frequency-domain features were obtained.

The other group of features used in training the prediction model consists of conventional image features that are computed from pixel information. Those features typically define color, texture and geometric features. Following Oliveira *et al.* (2018), 513 features in total were determined, among which 9 geometric variation features, 180 texture features generated using the gray level co-occurrence matrix (GLCM) method, 240 Haar-like features, 12 Hausdorff-based features, and 72 features from color spaces such as RGB, KSV, CIELAB, and CIELUV. All features extracted are summarized in Table 2.

2.2.3. Feature selection based on the mean decrease in impurity. Feature importance by default is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree (Li *et al.*, 2019).

Let denote by $MDI(f)$ the mean decrease impurity of feature f . It assesses the importance of each feature by summing up the number of splits (across all trees) that involve the feature, adjusted based on the proportion of samples it divides. More specifically,

$$MDI(f) = \frac{1}{T} \sum_{t=1}^T \sum_{i \in N_t} p(i|t) \left(1 - \sum_{k=1}^K p(k|i)^2\right) I(f = f_i), \quad (2)$$

where T is the number of trees, N_t is the set of nodes in tree t , $p(i|t)$ is the proportion of samples reaching node i in tree t , K is the number of classes, $p(k|i)$ is the proportion of samples of class k at node i , f_i is the feature used to split node i , and $I(\cdot)$ is the indicator function.

The computed values of $MDI(f)$ are used to set the MDI-based thresholds in order to determine the importance of features for training the classification model.

2.2.4. Feature selection via SHAP values. The feature importance and selection was determined based on the sum of absolute SHAP values for feature k defined as

Table 2. Main groups of extracted features.

Group	Features
Graph nodes	local efficiency, local clustering coefficient, nodal strength, nodal betweenness centrality, closeness centrality, eccentricity
Entire graph	characteristic path length, global efficiency, global clustering coefficient, density, global assortativity
Frequency domain	vector of graph Fourier transform, energy, power, entropy, amplitude
Geometrical features	area, perimeter, equivalent diameter, compactness, circularity, solidity, rectangularity, aspect ratio, eccentricity
Color spaces	mean of pixel values, variance of pixel values, standard deviation of pixel values
Channels of color space	mean of image pixel values in channel, variance of image pixel values in channel, standard deviation of image pixel values in channel, min of image pixel values in channel, max of image pixel values in channel, skew of image pixel values in channel, Hausdorff dimension
GLCM features	contrast, dissimilarity, homogeneity, energy, correlation, variance, sum entropy, sum average, difference variance, difference entropy, maximal correlation coefficient, information measure of correlation
Haar wavelet	energy, entropy

$$\text{shapSum}_k = \sum_i \sum_j |s_{ijk}|. \quad (3)$$

Thus, shapSum_k is the sum of absolute SHAP values s_{ijk} for feature k across all samples and outputs. For a particular feature k and model f , the SHAP value is computing using the formula

$$s_k(f) = \sum_{S \subseteq K \setminus k} \frac{|S|!(|K| - |S| - 1)!}{|K|!} [f(S \cup k) - f(S)], \quad (4)$$

where K is the set of all features, S is a subset of K that does not include feature k , $|S|$ is the number of elements

in S , $|K|$ is the total number of features, $f(S \cup k)$ is the output of the model with features in S and feature k , $f(S)$ is the output of the model with features in S only.

2.3. Performance metrics. This subsection introduces the metrics used to determined the performance of the skin lesion prediction model, which is a multi-class classification model in the study.

Suppose that y defines the true label, while \hat{y} is the predicted label. The accuracy metrics is given as

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^n I(y_i = \hat{y}_i), \quad (5)$$

where I is the indicator function that returns 1 if y_i equals \hat{y}_i and 0 otherwise, n is the total number of predictions.

In general, the F1-score metric is the preferred choice in situations involving imbalanced class distribution, as it offers a balanced assessment that considers both precision and recall (Type I and Type II errors). In contrast to binary classification, a multi-class model produces individual F1-scores for each class. Therefore, the weighted F1-score is computed using the formula

$$F1_{\text{weighted-score}} = \sum_{i=1}^n \frac{N_i}{N} \cdot \frac{2 \cdot \frac{TP_i}{TP_i + FP_i} \cdot \frac{TP_i}{TP_i + FN_i}}{\frac{TP_i}{TP_i + FP_i} + \frac{TP_i}{TP_i + FN_i}}, \quad (6)$$

where w_i is the weight of the i -th label (the number of samples from the i -th label divided by the total number of samples), N_i is the number of samples from the i -th label, and N is the total number of samples. As could be seen from this formula, the weighted F1-score is the average taking into account the proportion for each label in the sample.

3. Experimental results

Figure 3 displays the examples of constructed superpixel graphs for all types of skin lesions considered in the paper. It could be seen that vertices of the superpixel graph tend to cluster closely together in darker regions of skin lesion, as opposed to brighter ones.

3.1. Binary skin lesion classification. Table 3 presents 5-fold cross validation results of a binary skin lesion classification in terms of accuracy and weighted F1-score. Having in mind that the image set is comparatively small, the model performance results are in line with our expectations. It may be concluded that the highest accuracy and the best weighted F1-score were obtained by selecting features based on the threshold applied for shapSum values. However, the performance metrics declined in situations where either all features were employed or features were selected based on the threshold established for MDI values.

Figure 4 shows the discriminatory power of the binary classification model for various shapSum-based thresholds. The model is more prone to making errors when predicting Class 1, but the overall accuracy is comparatively high. Notably, the best performance was achieved for shapSum thresholds close to 1, which is better than setting a lower value for the bound of SHAP values. Figure 5 suggests that the use of MDI values for feature selection with larger thresholds negatively influences the prediction results, as the discrimination between two classes is worse. On the other hand, setting the threshold of MDI comparatively low allows to employ more information in training model, however, still not enough in comparison to the results observed when shapSum-based thresholds were set for selecting features.

Next, we investigate how the model performance is sensitive to changes in the threshold value (see Figs. 6 and 7). It could be seen that weighted F1-score varies slightly for different shapSum values, which we consider as an advantage. On the contrary, weighted F1-score is very sensitivity to the adjustments of MDI-based thresholds, potentially affecting the stability of model performance.

3.2. 5-Class skin lesion prediction. Table 4 summarizes the 5-fold cross validation results of 5-class skin lesion prediction in terms of accuracy and weighted F1-score. In comparison with binary classification of skin lesion (see Table 3), the performance metrics are getting worse, which is not surprising, as the prediction difficulty grows with an increase in the number of classes. From Table 4, it is evident that the highest accuracy and highest weighted F1-score were achieved through the selection of features via SHAP values. Particularly noteworthy is the substantial improvement in the weighted F1-score when features were chosen using SHAP values compared with the scenario where all features were used. It is conceivable that, due to redundancy or an overly extensive list of features, performance in predictions is getting worse.

In Fig. 8, the diagonal elements indicate a very good performance of the developed model in predicting the type of skin lesion. Notably, the model performs best in predicting Class 2, which is a melanoma. However, the model encounters challenges when predicting Class 1, which is mainly confused with Class 2 and less often with Class 3. The adjustment of the upper limit of SHAP values suggests that the setting of an excessively high threshold may not be optimal. More specifically, when the threshold is too high, only a small number of features meet the criteria, which is not enough to predict the type of skin lesion.

Figure 9 suggests that in the case when the features were selected using the threshold based on the MDI value the misclassification of diagnosis had a tendency to increase when compared with results observed in Fig. 8.

Table 3. 5-Fold cross validation results for a binary case.

Performance metric	All features	MDI-based features	SHAP-based features
F1-score	0.9225	0.9478	0.9667
Weighted F1-score	0.8312	0.8530	0.9248

Table 4. 5-Fold cross validation results for 5-class prediction.

Performance metric	All features	MDI-based features	SHAP-based features
F1-score	0.6818	0.6970	0.7213
Weighted F1-score	0.6366	0.6818	0.7121

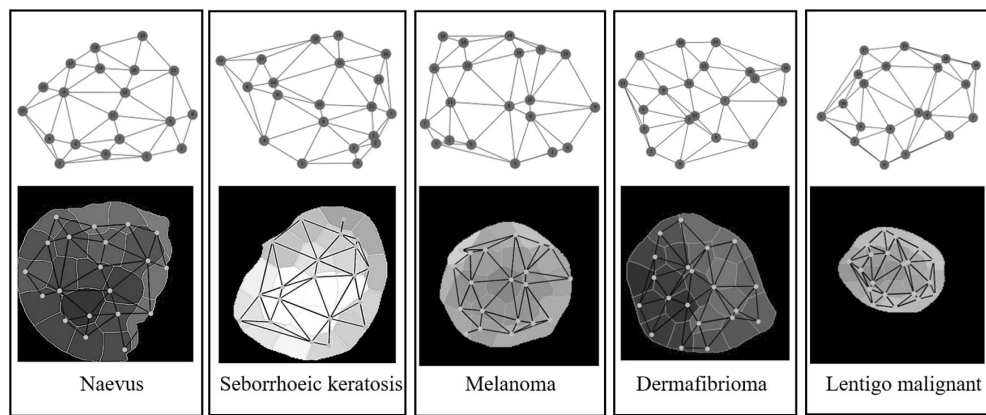


Fig. 3. Generated Delaunay triangulation of the graph for instances of each class.

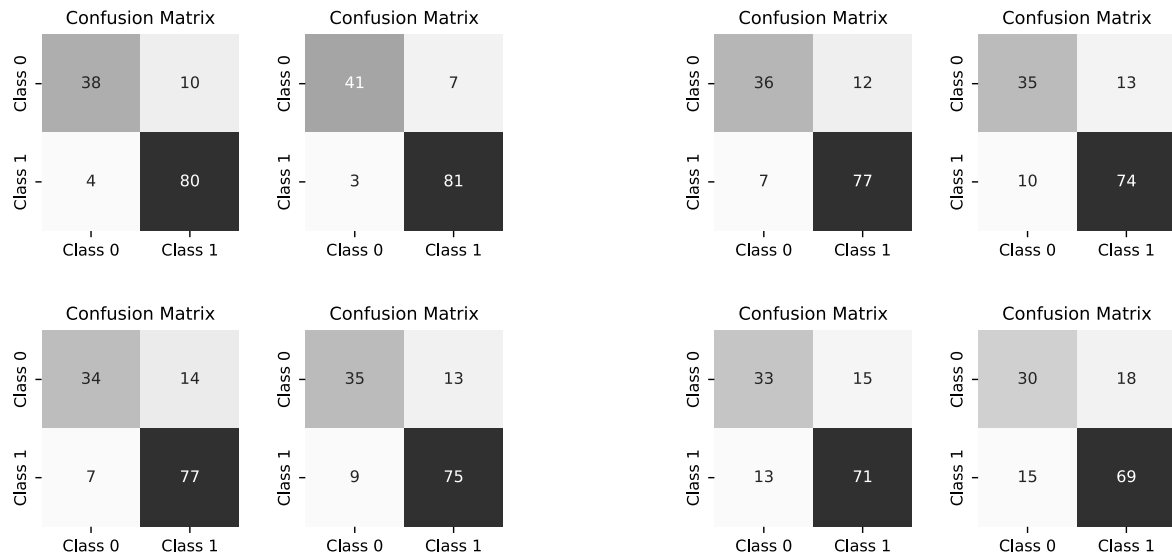


Fig. 4. Confusion matrices of binary classification using various shapSum-based thresholds: SHAP values ≥ 0.6690 (a), SHAP values ≥ 1.0028 (b), SHAP values ≥ 1.5259 (c), SHAP values ≥ 3.128 (d).

Fig. 5. Confusion matrices of binary classification using various MDI-based thresholds: MDI values ≥ 0.003 (a), MDI values ≥ 0.004 (b), MDI values ≥ 0.005 (c), MDI values ≥ 0.01 (d).

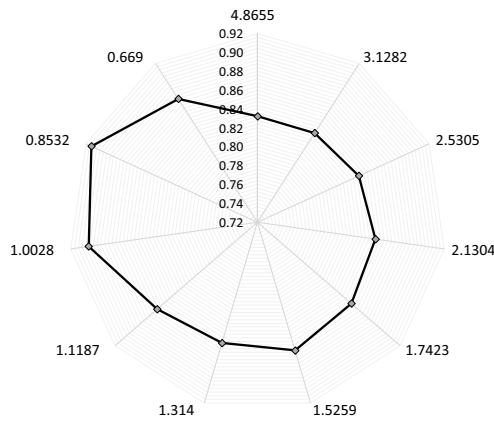


Fig. 6. Weighted F1-scores of 2-class classification using various shapSum based thresholds.

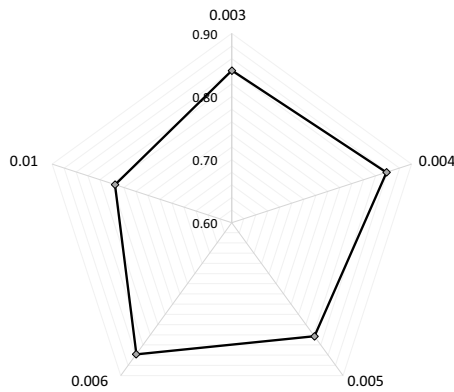


Fig. 7. Weighted F1-scores of 2-class classification using various MDI-based thresholds.

Setting an excessively large threshold for MDI makes the predictions less accurate, which may be influenced by an excessively low number of features needed for a good model performance.

Similarly to the case of binary classification, we explore the model stability in terms of weighted F1-score against the change in the threshold for shapSum values (see Fig. 10) and MDI values (see Fig. 11). The same tendency could be seen here as well, i.e., the weighted F1-score exhibits a higher degree of sensitivity to the change in the MDI value than in the scenario when shapSum values were used.

Table 5 displays the impact of superpixel graph-based features used in training both binary prediction and 5-class prediction models. The importance of features was determined by shape-based thresholds only. As shown in the table, the model performance significantly benefits from the the use of graph-based

Confusion Matrix

True labels \ Predicted labels	Class 0	Class 1	class 2	Class 3	Class 4
Class 0	7	0	2	1	2
Class 1	0	1	11	6	0
class 2	0	3	63	4	0
Class 3	0	1	5	12	0
Class 4	2	0	3	2	7

Confusion Matrix

True labels \ Predicted labels	Class 0	Class 1	class 2	Class 3	Class 4
Class 0	8	0	2	1	1
Class 1	0	4	8	5	1
True labels class 2	0	4	65	1	0
Class 3	0	1	4	13	0
Class 4	2	0	3	2	7

Confusion Matrix

True labels \ Predicted labels	Class 0	Class 1	class 2	Class 3	Class 4
Class 0	8	1	1	1	1
Class 1	0	2	10	6	0
True labels class 2	0	5	62	3	0
Class 3	0	0	5	13	0
Class 4	2	0	2	2	8

Confusion Matrix

True labels \ Predicted labels	Class 0	Class 1	class 2	Class 3	Class 4
Class 0	7	0	2	1	2
Class 1	0	4	9	5	0
True labels class 2	1	4	61	4	0
Class 3	0	1	4	13	0
Class 4	3	0	3	2	6

Fig. 8. Confusion matrices for a 5-class classification task using various shapSum-based thresholds: SHAP values ≥ 0.6690 (a), SHAP values ≥ 1.0028 (b), SHAP values ≥ 1.5259 (c), SHAP values ≥ 3.128 (d).

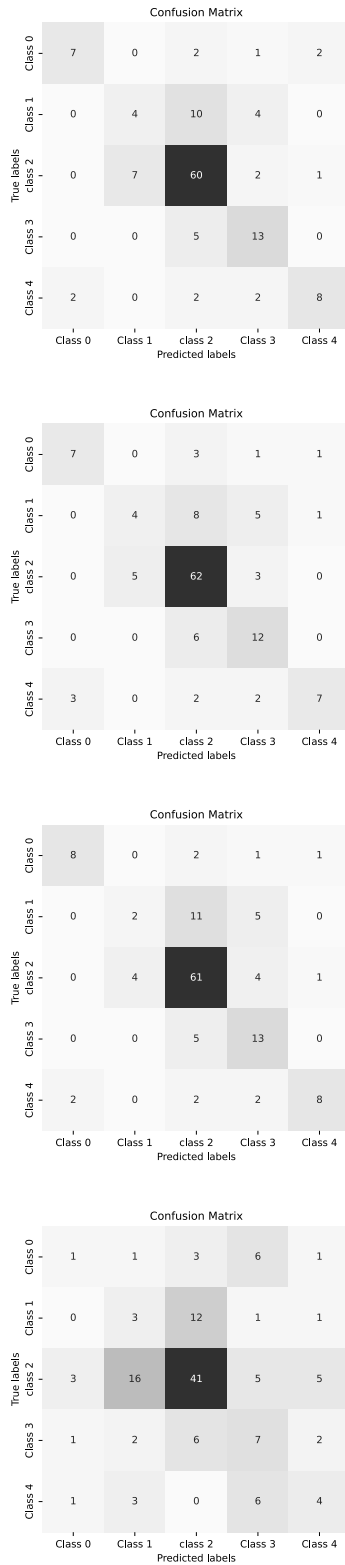


Fig. 9. Confusion matrices for a 5-class classification task using various MDI-based thresholds: MDI values ≥ 0.003 (a), MDI values ≥ 0.004 (b), MDI values ≥ 0.005 (c), MDI values ≥ 0.01 (d).

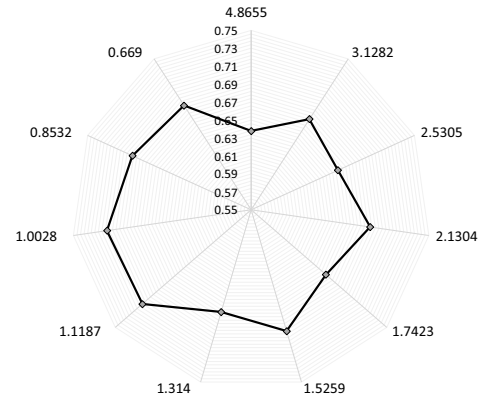


Fig. 10. Weighted F1-scores of 5-class classification using various shapSum-based thresholds.

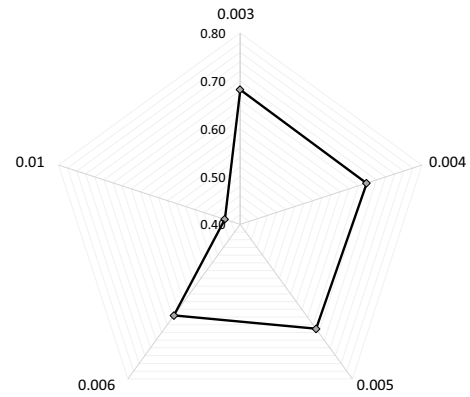


Fig. 11. Weighted F1-scores of 5-class classification using various MDI-based thresholds.

features, as the accuracy was improved by 1.11 (binary) and 1.13 (multi-class), while the weighted F1-score changed by 1.10 and 1.16, respectively.

In addition, the experiments carried out by varying the percentages of the testing and training samples showed that the best results were obtained with a testing sample between 25% and 40% for the 5-class classification task irrespective of the choice of criteria for the threshold (SHAP or MDI) for feature selection.

Although the field of computer-aided systems for skin lesion diagnosis is rapidly advancing with extensive research utilizing diverse machine learning methods, including deep learning architectures (Bibi *et al.*, 2023; Aldhyani *et al.*, 2022; Shetty *et al.*, 2022; Khan *et al.*, 2022) direct comparison with other studies is limited due to the unique nature of our non-public dataset. We can, however, draw comparisons with studies utilizing datasets with similar characteristics, including the number

Table 5. Impact of features set based on SHAP-based thresholds for the model performance.

	Binary classification		5-Class classification	
	Conventional features	Conventional + graph-based features	Conventional features	Conventional + graph-based features
Accuracy	0.8439	0.9375	0.6390	0.7213
F1	0.8437	0.9666	0.6271	0.7120
Weighted F1	0.8405	0.9248	0.6133	0.7121
Recall	0.8182	0.9687	0.6379	0.7213
Precision	0.8710	0.9645	0.6168	0.7030
Specificity	0.8788	0.9375	0.9076	0.8501

of classes, types of skin lesions, as well as the number of instances. Based on the most recent comprehensive review paper by Kassem *et al.* (2021), our binary classification accuracy aligns with previously reported ranges from 73.8% to 99.9%, while in the case of multi-class (5 classes) the accuracy value is at least 66.2%. These findings suggest that more data are needed in those classes where there are few instances and which perform the worst in terms of classification. Furthermore, further research should be carried out to investigate methods that can improve the classification performance.

4. Conclusions

This study delved into the realm of binary and multi-class classification of skin lesions. As the diverse set of features play an important role in the training machine learning model, the particular emphasis was placed on identifying essential features by establishing thresholds based on both MDI and SHAP values. In total, a set of 662 features was generated, which was composed of 513 image conventional and 149 graph-based features derived from superpixel graph constructed by Delaunay triangulation.

In general, the use of a threshold to determine feature importance and their contributions to classifier training played an essential role. First, employing all features resulted in lower accuracy and weighted F1-scores, indicating potential redundancy or an excessively large set of features. Second, feature selection based on the shapSum-based threshold outperformed the mean decrease in the impurity based threshold, which is true in the case of both binary and multi-class prediction models. Third, the performance metrics deteriorated considerably after the switch from a binary to a 5-class prediction model, which is not surprising, as the prediction difficulty enlarged remarkably. Fourth, it is evident that performance metrics, namely the weighted F1-score, was more sensitive to changes in the threshold

of the MDI-based value than SHAP values, suggesting a greater stability of model performance when SHAP values were used for feature selection. And, finally, graph-based features played a substantial role in enhancing the prediction model, as evidenced by an improvement of at least 1.10 in accuracy and weighted F1-score.

The direct comparison with other studies is limited due to the unique nature of our non-public dataset. We can, however, draw comparisons with studies utilizing datasets with similar characteristics, including the number of classes, types of skin lesions, as well as the number of instances. Based on the most recent comprehensive review paper (Aloupogianni *et al.*, 2022), our binary classification accuracy aligns with previously reported ranges from 73.8% to 99.9%, while in the case of multi-class classification (5 classes) the accuracy value is at least 66.2%. These findings suggest that more data are needed in those classes, particularly those that performed poorly in terms of performance measures.

Further investigation can focus on the calibration of SHAP threshold values and the development of graphical explanations of feature importance through Explainable AI (XAI) methodologies, by merging SHAP values with deep learning models. Such a model enables the calculation of input importance relative to a reference by backpropagating contribution scores through the neural network and offers computational efficiency through an approximation to Shapley values. With such an approach, we can provide not only precise classification or segmentation models but also ensure valuable and trustworthy assistance for dermatologists. Further research also need to be carried out in order to fully automate the diagnosis of skin lesions, particularly, focusing on the removal of artefacts such as hair, veins, surgical markings, light reflections, etc. This adds additional complexity to this task, which requires an separate research (Winkler *et al.*, 2019). This difficulty is also experienced due to the various types of images available.

Acknowledgment

This research was made under the European Union's Horizon Europe programme under the grant agreement no. 101059903, *Centre of Excellence for Sustainable Living and Working (SustAInLivWork)*.

References

- Abbes, W., Sellami, D., Marc-Zwecker, S. and Zanni-Merk, C. (2021). Fuzzy decision ontology for melanoma diagnosis using KNN classifier, *Multimedia Tools and Applications* **80**: 25517–25538.
- Abdelhalim, I.S.A., Mohamed, M.F. and Mahdy, Y.B. (2021). Data augmentation for skin lesion using self-attention based progressive generative adversarial network, *Expert Systems with Applications* **165**: 113922.
- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S. (2010). Slic superpixels, *Technical Report 149300*, Ecole Polytechnique Fédérale de Lausanne, Lausanne.
- Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S. and Islam, M. (2022). An effective skin cancer classification mechanism via medical vision transformer, *Sensors* **22**(11): 4008.
- Aldhyani, T.H.H., Verma, A., Al-Adhaileh, M.H. and Koundal, D. (2022). Multi-class skin lesion classification using a lightweight dynamic kernel deep-learning-based convolutional neural network, *Diagnostics* **12**(9), Article no. 2048.
- Ali, K., Shaikh, Z.A., Khan, A.A. and Laghari, A.A. (2022). Multiclass skin cancer classification using EfficientNets—A first step towards preventing skin cancer, *Neuroscience Informatics* **2**(4): 100034.
- Aloupogianni, E., Ishikawa, M., Kobayashi, N. and Obi, T. (2022). Hyperspectral and multispectral image processing for gross-level tumor detection in skin lesions: A systematic review, *Journal of Biomedical Optics* **27**(06): 060901.
- Alwakid, G., Gouda, W., Humayun, M. and Sama, N.U. (2022). Melanoma detection using deep learning-based classifications, *Healthcare* **10**(12): 2481.
- Annaby, M.H., Elwer, A.M., Rushdi, M.A. and Rasmy, M.E.M. (2021). Melanoma detection using spatial and spectral analysis on superpixel graphs, *Journal of Digital Imaging* **34**(1): 162–181.
- Araújo, R.L., Araújo, F.H.D.d. and Silva, R.R.V.e. (2021). Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning, *Multimedia Systems* **28**(4): 1239–1250.
- Ashraf, H., Waris, A., Ghafoor, M.F., Gilani, S.O. and Niazi, I.K. (2022). Melanoma segmentation using deep learning with test-time augmentations and conditional random fields, *Scientific Reports* **12**(1): 3948.
- Bibi, S., Khan, M.A., Shah, J.H., Damaševičius, R., Alasiry, A., Marzougui, M., Alhaisoni, M. and Masood, A. (2023). MSRNet: Multiclass skin lesion recognition using additional residual block based fine-tuned deep models information fusion and best feature selection, *Diagnostics* **13**(19), Article no. 3063.
- Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H. and Halpern, A. (2017). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC), *2018 IEEE 15th International Symposium on Biomedical Imaging, Washington, USA*, pp. 168–172.
- Combaila, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S. and Malvey, J. (2019). BCN20000: Dermoscopic lesions in the wild, *arXiv*: 1908.02288.
- Damian, F.-A., Moldovanu, S. and Moraru, L. (2022). Melanoma detection using a random forest algorithm, *2022 E-Health and Bioengineering Conference (EHB), Iasi, Romania*.
- Dildar, M., Akram, S., Irfan, M., Khan, H.U., Ramzan, M., Mahmood, A.R., Alsaiani, S.A., Saeed, A.H.M., Alraddadi, M.O. and Mahnashi, M.H. (2021). Skin cancer detection: A review using deep learning techniques, *International Journal of Environmental Research and Public Health* **18**(10): 5479.
- Emery, J.D., Hunter, J., Hall, P.N., Watson, A.J., Moncrieff, M. and Walter, F.M. (2010). Accuracy of siascopy for pigmented skin lesions encountered in primary care: Development and validation of a new diagnostic algorithm, *BMC Dermatology* **10**(1): 9.
- Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N. and Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC), *arXiv*: 1605.01397.
- Hurtado, J. and Reales, F. (2021). A machine learning approach for the recognition of melanoma skin cancer on macroscopic images, *TELKOMNIKA (Telecommunication Computing Electronics and Control)* **19**(4): 1357.
- Ilişanu, M.-A., Moldoveanu, F. and Moldoveanu, A. (2023). Multispectral imaging for skin diseases assessment—State of the art and perspectives, *Sensors* **23**(8): 3888.
- Jocher, G., Chaurasia, A. and Qiu, J. (2023). Ultralytics YOLOv8, <https://github.com/ultralytics/ultralytics>.
- Kassem, M.A., Hosny, K.M., Damaševičius, R. and Eltoukhy, M.M. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review, *Diagnostics* **11**(8), Article no. 1390.
- Khan, M.A., Sharif, M.I., Raza, M., Anjum, A., Saba, T. and Shad, S.A. (2022). Skin lesion segmentation and classification: A unified framework of deep neural network features fusion and selection, *Expert Systems* **39**(7): e12497.

- Li, X., Wang, Y., Basu, S., Kumbier, K. and Yu, B. (2019). A debiased MDI feature importance measure for random forests, *Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada*, pp. 8049–8080.
- Liutkus, J., Kriukas, A., Stragyte, D., Mazeika, E., Raudonis, V., Galetzka, W., Stang, A. and Valiukeviciene, S. (2023). Accuracy of a smartphone-based artificial intelligence application for classification of melanomas, melanocytic nevi, and seborrheic keratoses, *Diagnostics* **13**(13): 2139.
- Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S. and Rozeira, J. (2013). Ph2—A dermoscopic image database for research and benchmarking, *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan*, pp. 5437–5440.
- Murugan, A., Nair, S.H. and Kumar, K.P.S. (2019). Detection of skin cancer using SVM, random forest and KNN classifiers, *Journal of Medical Systems* **43**(8): 1–9, DOI: 10.1007/s10916-019-1400-8.
- Navarrete-Dechent, C., Dusza, S.W., Liopyris, K., Marghoob, A.A., Halpern, A.C. and Marchetti, M.A. (2018). Automated dermatological diagnosis: Hype or reality?, *Journal of Investigative Dermatology* **138**(10): 2277–2279, DOI: 10.1016/j.jid.2018.04.040.
- Oliveira, R.B., Pereira, A.S. and Tavares, J.M.R. (2017). Skin lesion computational diagnosis of dermoscopic images: Ensemble models based on input feature manipulation, *Computer Methods and Programs in Biomedicine* **149**: 43–53, DOI: 10.1016/j.cmpb.2017.07.009.
- Oliveira, R.B., Pereira, A.S. and Tavares, J.M.R.S. (2018). Computational diagnosis of skin lesions from dermoscopic images using combined features, *Neural Computing and Applications* **31**(10): 6091–6111.
- Oukil, S., Kasmi, R., Mokrani, K. and García-Zapirain, B. (2021). Automatic segmentation and melanoma detection based on color and texture features in dermoscopic images, *Skin Research and Technology* **28**(2): 203–211.
- Rashid, H., Tanveer, M.A. and Aqeel Khan, H. (2019). Skin lesion classification using GAN based data augmentation, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany*, pp. 916–919.
- Rey-Barroso, L., Burgos-Fernández, F., Delpueyo, X., Ares, M., Royo, S., Malveyh, J., Puig, S. and Vilaseca, M. (2018). Visible and extended near-infrared multispectral imaging for skin cancer diagnosis, *Sensors* **18**(5): 1441.
- SIIM-ISIC (2020). *International Skin Imaging Collaboration 2020 Challenge Dataset*, <https://challenge2020.isic-archive.com/>.
- Sharma, G., Wu, W. and Dalal, E.N. (2004). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations, *Color Research & Application* **30**(1): 21–30.
- Shetty, B., Fernandes, R., Rodrigues, A.P., Chengoden, R., Bhattacharya, S. and Lakshmana, K. (2022). Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, *Scientific Reports* **12**(1): 18134.
- Sunarya, C.A., Siswanto, J.V., Cam, G.S. and Kurniadi, F.I. (2023). Skin cancer classification using Delaunay triangulation and graph convolutional network, *International Journal of Advanced Computer Science and Applications* **14**(6), DOI: 10.14569/IJACSA.2023.0140685.
- Surówka, G. and Ogorzałek, M. (2022). Segmentation of the melanoma lesion and its border, *International Journal of Applied Mathematics and Computer Science* **32**(4): 683–699, DOI: 10.34768/amcs-2022-0047.
- Tembhurne, J.V., Hebbar, N., Patil, H.Y. and Diwan, T. (2023). Skin cancer detection using ensemble of machine learning and deep learning techniques, *Multimedia Tools and Applications* **82**(18): 27501–27524.
- Tschandl, P., Rosendahl, C. and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Scientific Data* **5**(1): 180161.
- Vikas Reddy, K. and Rama Parvathy, L. (2022). Accurate detection and classification of melanoma skin cancer using decision tree algorithm over CNN, in D.J. Hemanth et al. (Eds), *Advances in Parallel Computing Algorithms, Tools and Paradigms*, IOS Press, Amsterdam, pp. 321–326.
- Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y.M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *arXiv*: 2207.02696.
- Wen, D., Khan, S.M., Ji Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A.K., Liu, X. and Matin, R.N. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review, *The Lancet Digital Health* **4**(1): e64–e74.
- Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W. and Haenssle, H.A. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition, *JAMA Dermatology* **155**(10): 1135–1141.
- Zafar, M., Sharif, M.I., Sharif, M.I., Kadry, S., Bukhari, S.A.C. and Rauf, H.T. (2023). Skin lesion analysis and cancer detection based on machine/deep learning techniques: A comprehensive survey, *Life* **13**(1): 146.



Agne Paulauskaite-Taraseviciene is the head of the Centre for Artificial Intelligence and a professor at the Faculty of Informatics of the Kaunas University of Technology. Over the past 15 years, all the research projects in which she has been involved have focused on the field of AI. She has gained experience in the implementation and development of AI solutions in (bio)medicine, industry, and agriculture sectors, addressing tasks such as anomaly detection, forecasting, and the

analysis of complex images.



Kristina Sutiene holds an MSc in applied mathematics and a PhD in informatics from the Kaunas University of Technology. She is interested in mathematical modeling and analysis of stochastic, medical and economic systems, data analysis and anomaly detection, development of predictive and statistical models, and risk management in business systems.



Skaidra Valiukeviciene is the head of the Department of Skin and Venereal Diseases and a professor at the Lithuanian University of Health Sciences in Kaunas. The main research activities include investigations of skin biophysics related with healthy ageing using technologically modified or new created devices such as ultrasound, infrared thermography, spectrophotometry and optical images for non-invasive diagnostics of skin cancer and diabetes.



Nojus Dimsa holds a BSc degree in informatics from the Kaunas University of Technology and is currently pursuing an MSc degree in informatics of artificial intelligence at the same institution. With three years of hands-on experience, he has strong background in photogrammetry, medical image processing and computer vision domains.

Received: 15 February 2024

Revised: 19 May 2024

Re-revised: 2 July 2024

Accepted: 23 August 2024