

A HOLISTIC APPROACH TO MULTI-MODAL SKIN LESION DIAGNOSIS SUPPORTED BY STATISTICAL AND EXPLAINABILITY-BASED INVESTIGATION OF ARTIFACTS

Jakub Buler¹, Rafał Buler¹, Krystian Brzozowski^{1,2},
Maria Ferlin¹, Maciej Bobowicz^{1,2}, Michał Grochowski^{1,2,*}

¹*Gdańsk University of Technology,
Department of Intelligent Control and Decision Support Systems,
Narutowicza 11/12, 80-233 Gdańsk, Poland*

²*Medical University of Gdańsk,
2nd Division of Radiology,
Marii Skłodowskiej-Curie 3a, 80-210 Gdańsk, Poland*

**E-mail: michal.grochowski@pg.edu.pl*

Submitted: 19th July 2025; Accepted: 11th May 2026

Abstract

Skin cancer is a prevalent and potentially life-threatening condition, where early and accurate detection is crucial for effective treatment. Traditional diagnosis, based on visual assessment, faces challenges due to inter-class variability and the similarity between benign and malignant lesions. Although automated diagnostic systems are gaining attention, few offer comprehensive evaluations. This study addresses this gap by systematically assessing such systems and reporting extensive performance metrics.

This study contributes to both artificial intelligence and dermatological image analysis by introducing an automated system for multi-class skin lesion classification. The system was evaluated under various configurations, including original and artifact-inpainted images. Three approaches were employed: deep learning, radiomics-based and hybrid. Additionally, to enhance the interpretability, the system generates segmentation masks and attribution maps.

Models were evaluated using five-fold cross-validation on the International Skin Imaging Collaboration (ISIC) 2018 Challenge Task 3 training dataset, with final performance assessed on the official test set. The best-performing model achieved a micro accuracy of 79.76 ± 0.73 , macro F1-score of 73.62 ± 1.38 , precision of 74.80 ± 0.84 , and recall of 73.93 ± 2.19 , surpassing comparable methods reported in the literature.

The contribution to the research community is the release of a novel set of 57630 binary artifact labels for the ISIC-2018 Task 3 dataset, enabling research on the impact of visual artifacts such as hair, rulers, and interface fluids. This contribution is supported by statistical analysis and attribution map assessment, revealing significant differences in artifact occurrence across lesion classes and highlighting the utility of these labels in identifying and mitigating bias in skin lesion classification.

Keywords: radiomics, skin lesion classification, medical data analysis, explainable artificial intelligence, multi-modal data

1 Introduction

Skin cancers are recognized as one of the most prevalent cancers worldwide, ranking fifth among all types of cancer [1]. They constitute a broad and variable group of cancers developing from skin cells containing dark pigment melanin (malignant melanoma) or cells without it (e.g. basal cell carcinoma, etc.). Despite varying prognoses and risks of metastasis, all types can significantly affect physical and mental well-being. Risk factors for developing skin cancers include increased exposure to ultraviolet radiation or sunlight, ozone depletion, increased participation in outdoor activities, genetic predisposition and immunosuppression [2]. Preventive measures, such as sun protection and regular skin checks, are vital in reducing the risk of developing skin cancer.

Early detection is critical for effective treatment and improved health outcomes. The rising incidence of skin cancers [3] poses significant health challenges, affecting people of various demographics. An effective way to diagnose skin cancer is through dermatoscopy, which commonly employs such criteria as the ABCDE criteria [4], the 7-point checklist or the Menzies method [5], and the CASH (color, architecture, symmetry, and homogeneity) algorithm [6] to discriminate malignant melanoma from other types of skin cancers. Among these, the ABCDE rule developed by Kopf, Friedman, and Rigel (1985) is the most frequently employed in the diagnosis of melanoma. The diagnostic final score is derived from the assessment of four features: asymmetry (A), border irregularity (B), color variation (C), and differential structures (D). An additional criterion, Evolution (E), is often included to assess changes in the lesion over a few months. Omission of these methods and dependence on the naked-eye assessment carries a high risk of misdiagnosis or under-diagnosis of malignant melanoma of the skin.

The results of recent studies indicate that artificial intelligence (AI) based systems can reach almost human-level performance in many tasks, including medical data analysis for disease diagnosis. The widespread adoption of medical decision support systems for diagnosis, also known as Computer-Aided Diagnosis (CADx), has the potential to assist clinicians in making a diagnosis. Ma-

chine learning (ML) and AI-based CADx systems facilitate fast, accessible, even remote, objective, and cost-effective support in diagnosing, e.g. for screening purposes or as supplementary diagnostic assistance. A family of AI methods naturally predisposed to efficiently analyze large amounts of medical data includes deep neural networks (DNN).

Despite significant achievements which DNN have reached in many fields, DNN and accompanying learning algorithms still have many challenges to tackle. In the case of medical applications, the most important factors are: limited amount of representative and good-quality data, dataset artifacts, class imbalance within the datasets, poor performance evaluation for AI systems, lack of transparency and interpretability in both, the design phase and the decision-making afterwards.

The lack of standardized reporting and evaluation protocols remains a significant challenge, limiting the comparability of results and the generalizability of models to real-world clinical applications. While prior studies have made significant progress in skin lesion classification, most rely on customized evaluation setups, lack explainability, or omit broader image quality issues, such as artifacts, that may affect model reliability. Moreover, many previous studies rely on a narrow set of performance metrics, typically limited only to accuracy, overlooking broader aspects of model performance.

This study revisits the issue by strictly adhering to the official ISIC-2018 dataset splits, avoiding any custom partitions to ensure fair and reproducible benchmarking. The ISIC 2018 dataset is a widely adopted and valuable benchmark for skin lesion classification, offering a diverse range of lesion types, and real-world variability. The performance of three distinct machine learning pipelines is evaluated for the classification of skin lesions into seven categories: actinic keratoses and intraepithelial carcinoma (*akiec*), basal cell carcinoma (*bcc*), benign keratosis-like lesions (*bkl*), dermatofibroma (*df*), melanoma (*mel*), melanocytic nevi (*nv*), and vascular lesions (*vasc*), using dermoscopic images.

In recent years, there has been a notable increase in the incorporation of radiomic analyses into the domain of AI-driven medical diagnostics. Radiomic features, defined as quantitative descriptors extracted from medical images, capture tex-

ture, shape, and intensity patterns. These features offer a rich source of complementary information, thus their utilization in medical image classification tasks has been increasingly recognized by experts as a powerful approach to enhance model interpretability and diagnostic accuracy. Motivated by this, the study presents a deep learning pipeline (DLP), a radiomic-based pipeline (RBP) that incorporates an automatic segmentation model, and a hybrid one (HP) which combines the strengths of the two previous approaches. The DLP uses convolutional neural networks (CNNs) to identify features from dermoscopic images, whereas the radiomics-based approach extracts detailed, clinically interpretable features from automatically segmented lesions. The HP enables joint information flow between modalities via feature-level integration and shared training objectives. This design allows the model to exploit both textural detail and global context, offering improved diagnostic performance. The system was designed assuming clinicians could use it as a decision-support tool. Therefore, one of the main objectives was to ensure that the end user could interpret the system. In this regard, the system prediction is accompanied by lesion segmentation and visual explanations of the predictions adequate to the model.

Our study contributes significantly in the form of new labels for the ISIC-2018 Task 3 dataset, both for training – HAM10000 (Human Against Machine with 10,000 Training Images) – and dedicated test dataset for this challenge [7, 8]. This involved identifying and flagging any artifact present in the image data. Afterwards, quantitative and qualitative analyses were conducted to assess the potential impact of these artifacts on the results. To address class imbalance, a stratified 5-fold cross-validation (CV) was used to ensure a reliable model performance assessment. The proposed classification and segmentation pipelines were trained on the HAM10000 dataset and evaluated on the dedicated ISIC-2018 test sets. The dual focus on segmentation and classification also aligns with practical medical workflows, enhancing the study's relevance.

The findings demonstrate that both the hybrid and deep learning pipelines achieved the state-of-the-art performance, outperforming similar methods reported in the literature. Furthermore, the

proposed skin cancer diagnosis system performs strongly on an independent dataset, indicating high generalizability. By integrating into the diagnosis pipelines, one of the most efficient methods of eXplainable Artificial Intelligence (XAI), the Integrated Gradients (IG) [9], into the diagnosis pipelines, the system offers interpretable predictions, highlighting the importance of both visual and radiomic features, along with segmentation masks.

The principal contribution of this paper can be expressed as follows:

- proposition and in-depth analysis of radiomic-based, deep learning-based, and hybrid multi-modal approaches to multi-class skin lesion classification,
- publication of a novel set of binary artifact labels for ISIC-2018 Task 3 Challenge datasets (57 630 new label entries),
- publication of segmentation masks for ISIC-2018 Task 3 test dataset,
- presentation of a comprehensive set of performance metrics,
- quantitative statistical analysis of the datasets for the occurrence of artifacts such as hair, ruler marks, interface fluid, vignettes, and other, and qualitative impact assessment on the model predictions supported by explainability techniques.

Together, these contributions address a key gap in the literature: the lack of standardized, interpretable, and artifact-aware evaluation of skin lesion classification models, grounded in official benchmarking protocols.

The structure of the paper is as follows: Section 2 provides a brief description of the related works. Section 3 covers the methodology employed in developing and validating the proposed system, including a detailed description of the datasets, implemented pipelines, model interpretability, and dataset analysis. Section 4 presents the results of the experiments, including quantitative performance metrics, and the results of the quantitative and qualitative analyses. Section 5 discusses the implications of the findings. Finally, Section 6 summarizes the paper, provides recommendations, and identifies directions for future research.

2 Related Work

In previous studies, the potential of machine learning and deep learning approaches has been demonstrated in the context of skin lesion diagnosis [10, 11, 12]. The study by [13] proposed two multi-class skin lesion classification methods: one based on a single custom-built Convolutional Neural Network and the other based on a combination of several CNNs in a one-versus-all scenario. The HAM10000 dataset was utilized and partitioned into training, validation, and test sets with a split ratio of 75/5/20. The authors claim that their system achieved an accuracy of 77% when using the CNN alone, which improved to 92.9% with the one-versus-all approach.

The authors in [14] further explored the fine-tuning of several CNN architectures and included model ensembling to enhance the performance. The HAM10000 dataset was split into training and validation subsets with a 90/10 split ratio. Ultimately, an accuracy of 93.20% was achieved for both individual models, ResNetXt101 and InceptionResNetV2, while the ensemble of these models yielded 92.83% accurate predictions.

The work presented by [15] demonstrated the use of the differential evolution algorithm DE-ANN for training the artificial neural network. The HAM10000 dataset was used, with 500 data samples separated for testing. Following image segmentation, a range of geometrical and texture features were extracted, including those derived from the RGB color space, color-level co-occurrence matrix, and local binary pattern. Using these features, the DE-ANN approach achieved an accuracy of 97.7% in the binary classification task.

In [16], the authors selected a subset of 700 images from the HAM10000 dataset, with 100 images per class, and split the data into training and test sets with an 80/20 ratio. Due to the observed imbalance in the data, which was identified as the primary reason for the necessity of data selection, the authors subsequently enlarged these sets using augmentations. Notably, the test data was also augmented. Many machine learning methods were applied, including Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Naive Bayes, K-Nearest Neighbour, and a custom CNN. Even though their work yielded a test accu-

racy of 86.43%, the authors presented their training accuracy of 95.18% for comparison with other works.

In [17], the authors conducted the analysis of data splits for training and testing within the HAM10000 dataset, using ratios of 70/30, 80/20 and 85/15. The incorporation of the Soft Attention blocks into several widely used deep neural networks, including VGG, ResNet, Inception ResNetv2 and DenseNet architectures, was proposed. The performance of these modified models was compared to the baseline models – without the Soft-Attention mechanism. The best results were observed for the 85/15 split ratio, achieving an accuracy of 93.4% with Inception ResNetv2 with Soft Attention. Nevertheless, the methodology used to obtain these results remains somewhat ambiguous, as only the training and test sets were employed. The models were trained with early stopping, yet it remains unclear whether the stopping criteria were based on the test set or the validation set was derived through other means. This raises concerns about the reliability of the reported performance.

Another approach presented by [18] involved the use of machine learning with hand-crafted features extracted according to the ABCD rule, which has been proven to be helpful in skin cancer diagnosis. First, masks of skin lesions were obtained through thresholding-based image segmentation. The HAM10000 and ISIC-2017 datasets were used with a 70/30 split ratio for training and testing, achieving an accuracy of 97.3%. However, there is no clear indication of which dataset this result refers to – whether it applies to the HAM10000 dataset, the ISIC-2017 dataset, or a combination of both. In addition, the authors mention that they manipulated the training and test data to remove outliers and irrelevant data yet provide no justification for these decisions. No information was provided on how the datasets were merged or handled, which raises significant concerns about the validity of the reported results.

Notably, the authors in [19] utilized the ISIC 2016–2020 datasets, applying exclusion criteria to eliminate duplicate entries, patients lacking demographic information, missing images, and other inconsistencies. Ultimately, the curated ISIC-2018 dataset was designated as the test set, while the remaining data constituted the training set. The

methodology involved the integration of a deep learning and radiomics-based approach, further enriched with patient metadata, to enhance skin lesion analysis. The hybrid model demonstrated strong performance across binary and multi-class classification tasks, achieving AUROC, sensitivity, and specificity ranging from 93.9% to 99%.

In the study [20] the authors have claimed to use the ISIC-2018 dataset, however, only the Task 3 training data part was taken (HAM10000 dataset), followed by selection of three classes from a total of seven. A split into training and test subsets with an 80/20 ratio was performed, followed by a further training and validation split. The proposed CNN ensemble aggregation method with a directed acyclic graph yielded a balanced accuracy of 76.6%. It is challenging to draw a valid comparison with other results due to how the subset of data was selected from the prepared training and test sets. Similarly, the work in [21] employs only three classes from the ISIC-2018 Task 3 training data (HAM10000). The training, validation and test sets were split with a 70/10/20 ratio. The authors proposed a framework that incorporates one multi-class CNN and six binary CNNs, which were modeled as players in a game theory context. EfficientNet-B5 was used as the backbone for this framework, and two distinct player strategies were implemented. The first strategy involved clustering the predictions by confidence level (i.e. confident, medium, or uncertain), while the second strategy focused on making the final prediction. The models' output probabilities determined the payoff. Their results showed an Area Under the Receiver Operating Characteristic Curve ranging from 0.93 to 0.97 for the three selected classes. The balanced accuracy varied from 81% to 94% across one multi-class task and six binary tasks, including one-vs-one and one-vs-all comparisons.

Ensemble methods have also been extensively explored for this type of task. For example, an ensemble integrating the features from 15 deep learning models achieved 80.3% balanced multi-class accuracy on a subset of the ISIC-2018 Task 3 test set [22]. In turn, [23] have combined models trained on different image sizes and image crops of HAM10000, resulting in 90 models that achieved a balanced accuracy of 86.2% on the ISIC-2018 Task 3 test set. While these approaches achieve excellent

performance metrics, their design often prioritizes accuracy over transparency. The resulting complexity of such ensemble systems poses a challenge in gaining the trust of medical professionals in decision support systems, who may be concerned about relying on poorly understood models. This highlights the importance of balancing performance and interpretability in systems intended for medical applications.

Currently, different variations of the original datasets are being used in the literature, which makes it challenging to directly compare the performance evaluations. Establishing consistent and comparable evaluation criteria is critical to advancing research in this area and enabling valid comparisons between the various methodologies. However, few studies have chosen to conduct research and report findings using the ISIC-2018 dataset in line with its original intent. Nevertheless, some notable studies have adhered to the dataset's intended purpose.

For instance, [24] proposed a hierarchical classification framework for skin lesions with three different taxonomic structures. The first structure, the two-level taxonomy, starts with identifying a lesion as melanocytic or non-melanocytic, followed by a differential diagnosis, and the three-level taxonomy adds an intermediate step for non-melanocytic lesions, classifying them as benign or malignant before the final diagnosis. On the other hand, the inverted taxonomy begins with malignancy determination, then distinguishes the melanocytic status, and ends with the differential diagnosis. A balanced accuracy of 72.6% for three-level taxonomy was achieved on the ISIC-2018 Task 3 test set with DenseNet-161 as the image encoder.

Moreover, [25] used merged ISIC 2017-2020 datasets and simultaneously proposed a strategy for removing duplicates to curate the dataset. The authors fine-tuned commonly used pre-trained DNNs and then tested them, among others, on the ISIC-2018 Task 3 test set, achieving the highest balanced accuracy of 62.1% with EfficientNet-B0.

Another study that utilized the training and test sets of the ISIC-2018 Task 3 dataset was conducted by [26]. First, the authors trained the ResNet-34 model using supervised learning, achieving an average accuracy of 77.8%. Moreover, the authors proposed an extended version of the system developed

through reinforcement learning. Their model was based on a deep Q-learning algorithm and used the multi-class probabilities previously learned by the supervised model. The advantage of their approach was due to the adaptation of the reward function in the form of a reward table provided by dermatologist experts. It allowed them to value correct and incorrect diagnoses depending on the class. For instance, classifying melanoma resulted in a higher reward and penalty for correct and incorrect categorizations compared to the classification of vascular lesions. The integration of RL into the system significantly improved sensitivity, increasing from 61.4% to 79.5% for melanoma and from 79.6% to 87.1% for basal cell carcinoma. In addition, the average accuracy increased to 79.2%.

The reviewed studies exhibit high variability in the data used, making direct comparison of results challenging. For a comprehensive understanding of the diversity of the data utilized and its training, validation, and test splits, refer to [25]. Table 1 provides a summary of the review of related work carried out for this study, which goes a little beyond what is discussed in [25]. Although a few of the studies allow valid comparisons to be made, they often lack comprehensive reporting of the full range of relevant measures. For instance, metrics such as accuracy and balanced accuracy are commonly reported. However, precision, recall, F1-score or specificity are frequently omitted, and when present, the aggregation method (e.g., macro or weighted average) is rarely specified. Moreover, the absence of class-wise performance metrics in reporting further complicates the understanding of the model's behavior across all categories. This is particularly important in unbalanced scenarios and for ensuring medical practicality, as certain classes, such as malignant melanoma, are of greater clinical importance due to a much worse prognosis.

3 Materials and Methods

3.1 Materials

This study utilized the ISIC-2018 Challenge Task 3 [27, 28] dataset, which comprises a dedicated training dataset referred to as HAM10000 [7, 8] and a dedicated test dataset used for external evaluation. The HAM10000 dataset was utilized to train both the classification and segmenta-

tion models, while the entirely held-out ISIC-2018 Task 3 test set was employed to evaluate classification performance. Additionally, the ISIC-2018 Task 1 test set was exclusively used for evaluating the segmentation model. It should be underlined that segmentation masks had to be generated beforehand for radiomic feature extraction from the ISIC-2018 Task 3 test set using a model developed on the HAM10000 dataset.

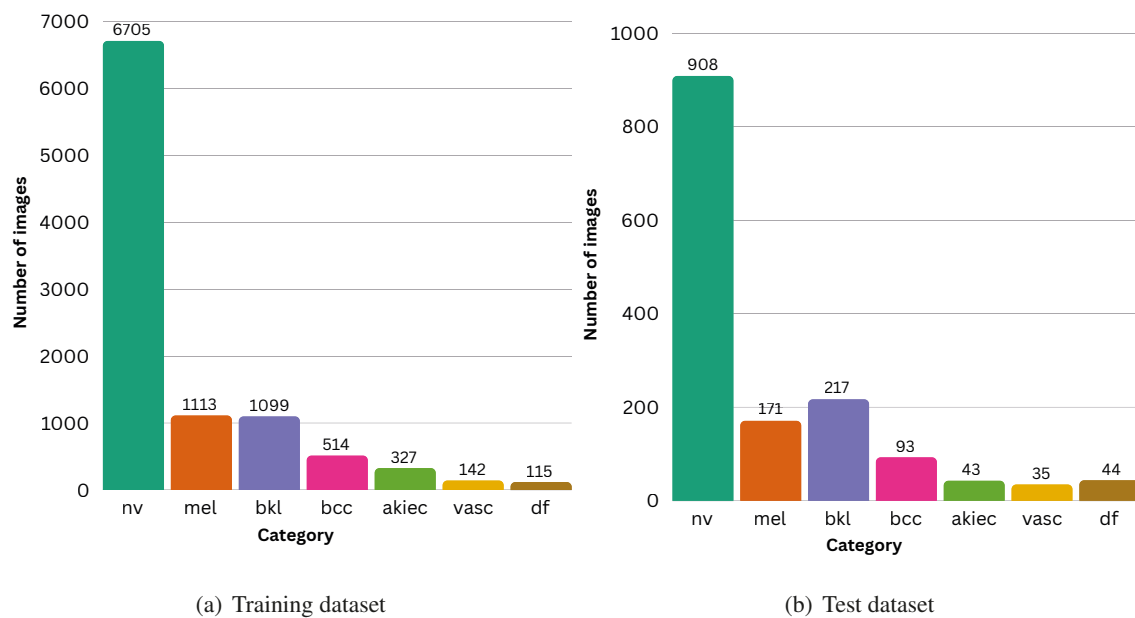
The HAM10000 consists of 10015 dermoscopic images labeled in seven classes, paired with binary segmentation masks, while the ISIC-2018 Task 3 test set consists of 1511 consistently categorized images without segmentation masks. This dataset is widely recognized as a benchmark due to its rich collection of dermoscopic images, each of size 450×600 px. It provides a diverse array of skin lesions for classification, detection, and segmentation tasks. There are seven categories of skin lesions: actinic keratoses and intraepithelial carcinoma (*akiec*), basal cell carcinoma (*bcc*), benign keratosis-like lesions (*bkl*), dermatofibroma (*df*), melanoma (*mel*), melanocytic nevi (*nv*), and vascular lesions (*vasc*). The class counts in the training and test sets are shown in Figure 1. It can be seen that there is a significant class imbalance, with the melanocytic nevi class being highly prevalent while others are comparatively under-represented. The aforementioned imbalance introduces challenges for model training and requires a comprehensive set of performance metrics to ensure reliable evaluation. Furthermore, the dataset comprises a diverse and representative set of real-world clinical cases from various medical institutions and research centers as part of the ISIC initiative. While this diversity is crucial for model reliability and generalization, it also adds complexity to the training process.

3.2 Methods

This section provides a description of the development process of the end-to-end diagnostic pipeline with interchangeable classifiers, comprising the following: data preprocessing, lesion segmentation, radiomic-based pipeline, deep learning-based pipeline or hybrid pipeline for classification, and post hoc model explanations. This framework is designed to be deployable in real-world medical applications, with each component contribut-

Table 1. Summary of studies on skin lesion classification based on the ISIC-2018 datasets.

STUDY	DATASETS	NO. OF CLASSES	NO. OF IMAGES	DATA SPLIT
[13]	HAM10000	7	10015	1-fold 75/5/20
[14]	HAM10000	7	10015	1-fold 90/10
[15]	HAM10000, PH2	2	10215	1-fold (1000 test samples)
[16]	HAM10000	7	700	1-fold 80/20
[17]	HAM10000	7	10015	70/30, 80/20, 85/15
[20]	HAM10000	3	8917	3-fold, 80/20
[21]	HAM10000	3	8917	1-fold 70/10/20
[22]	ISIC-2018 + other	7	30324	1-fold 90/10
[24]	ISIC-2018	7	10015 (1512)	5-fold (Dedicated Test)
[23]	ISIC-2016–18	7	12927 (1512)	5-fold (Dedicated Test)
[25]	ISIC-2017–20	7	9810 (1512)	Dedicated
[26]	ISIC-2018	7	10015 (1511)	20-fold (Dedicated Test)
[19]	ISIC-2016–20 + other	8	43534	5-fold

**Figure 1.** Class distributions in the ISIC-2018 Challenge Task 3 dataset.

ing critically to accurate and interpretable diagnostic outcomes. Additionally, this section provides a detailed description of a novel dataset analysis, including new labels identifying the presence of the features beyond the primary focus of interest, such as hair, ruler marks, interface fluid, and vignettes.

3.2.1 Data Preprocessing

Consistent image processing procedures were maintained in segmentation and classification tasks across DLP, RBP, and HP, including radiomics extraction in RBP and HP. The preprocessing phase included a number of sequential steps: a center square crop from 450×600 px to 450×450 px, followed by resizing to 224×224 px while preserving the aspect ratio. Hair removal was implemented using a simple thin contour morphological inpainting technique. The parameters used in the hair removal algorithm were selected based on their common application in community-driven solutions for ISIC datasets on Kaggle¹. In practice, these parameters have been shown to effectively suppress hair artifacts without noticeably distorting the lesion regions. The images were then standardized with ImageNet [29] statistics, aligning with the transfer learning approach employed. Morphological operations were applied to the segmentation masks to improve structural integrity by removing noise and filling in empty areas within the lesions. Subsequently, a connected component analysis was performed to retain only the largest lesion mask area, which typically corresponds to the primary lesion. Using masks prepared in such a way, training, validation and test set radiomic features were extracted and then standardized feature-wise using the training fold statistics.

The augmentation pipeline included various transformations such as flipping, cropping, rotating, color adjustments, affine transformations, and perspective distortions to obtain diverse data representations. Online augmentations, using continuous transformations and combinations of discrete ones, resulted in almost every sample being unique. Note that the degree of similarity to the original image remains very high. The test and validation sets were preprocessed as described in this section and excluded from the augmentation processes to ensure their independence and unbiased nature in

the final model evaluation. Due to the significant imbalance in the distribution of classes, the dataset was divided into five equal-sized folds for CV using a class-stratified sampling in HAM10000 proportions. In each iteration, one fold was held out as the validation set, while the remaining four folds were used for the training.

3.2.2 Lesion Segmentation

The model used for segmentation of lesions was of great importance for incorporating radiomic features into the classification of skin lesions within the RBP and HP. To address this, a lesion segmentation model was trained using the image-mask pairs from the training HAM10000 dataset, thereby enabling the generation of masks for the incomplete external ISIC-2018 Task 3 test set. In this study three distinct deep learning-based segmentation architectures: U-Net [30], U-Net++ [31], and DeepLabV3+ [32] were trained and carefully evaluated. Each architecture was equipped with an EfficientNet-B7 [33] as a backbone of the encoder and initialized with the ImageNet [29] pre-trained weights. The architecture that exhibited the optimum performance, U-Net, as determined by validation loss, was therefore selected for further utilization in subsequent downstream tasks.

3.2.3 Deep Learning Pipeline

The DLP pipeline (Figure 2) included the evaluation of four distinct CNN architectures to determine the most effective one. The evaluated backbones were: ResNet34 and ResNet50 [34], DenseNet121 [35], and EfficientNet-B0 [33]. Each model leveraged transfer learning by initializing with ImageNet [29] pre-trained weights. The last fully connected layer was replaced to align the models into a seven-class lesion classifier, and a softmax activation function was used to obtain the class probabilities. Among several deep neural network architectures evaluated during preliminary experiments, the model achieving the best performance on the validation set was selected for subsequent analyses – EfficientNet-B0. As these exploratory comparisons were not central to the study's objectives, detailed results are omitted. Once its head is detached, the selected architecture will act as a feature extractor for the hybrid framework.

¹<https://www.kaggle.com/code/vatsalparsaniya/melanoma-hair-remove>

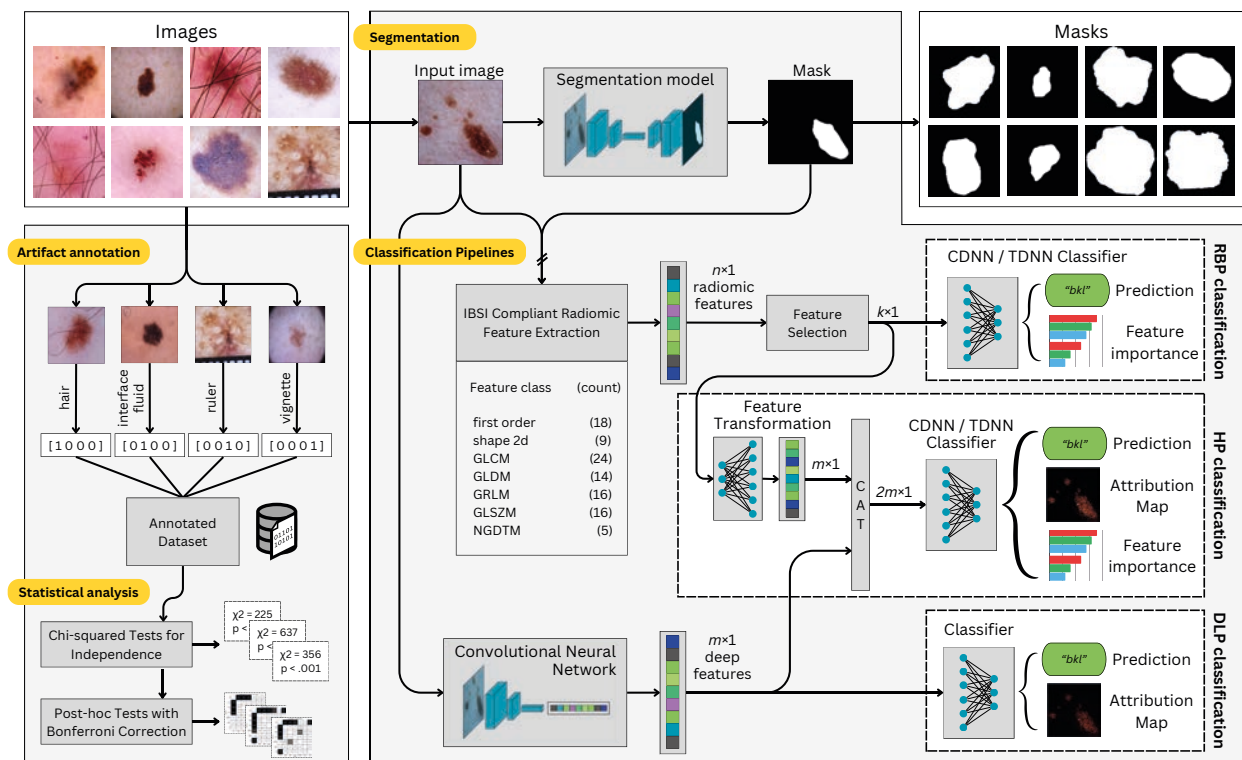


Figure 2. Illustration of the methodology including segmentation pipeline, classification pipelines, manual artifact annotation, and subsequent statistical analysis. Feature Selection block includes duplicate removal and LASSO; n – count of original radiomic features for RGB and grayscale image representations; k – count of radiomic features remaining after feature selection; m – deep feature vector size, varying with the model architecture; CAT – Feature Concatenation; CDNN – Custom Deep Neural Network; TDNN – Tabular Deep Neural Network.

To get insight into the model's underlying decision-making process, the local axiomatic visual explanation method, Integrated Gradients (IG) [9], was employed to provide fine-grained attribution maps. Various baselines were evaluated for Integrated Gradients, including zero image baseline, uniform image, and blurred image. Among these, the uniform image baseline was the most appropriate, as it provides an universally effective contrast for darker and lighter image regions, such as hair or ruler marks. The Grad-CAM method [36] was also analyzed; however, its visualizations were overly coarse, as the lesions often spanned nearly the entire image, resulting in a limited ability to provide precise heat maps facilitating the identification of image regions that are most important for the prediction.

3.2.4 Radiomic-based Pipeline

The RBP pipeline (Figure 2) incorporates the segmentation model selected from the earlier evaluation of the three architectures (Section 3.2.2). Before feature extraction, the images were preprocessed as described in Section 3.2.1. Subsequently, the images and their associated masks were utilized to compute the radiomic features. Radiomics are quantitative descriptors extracted from medical images that capture, among others, the texture, shape, intensity, and spatial distribution of the pixel values within the region of interest [37]. It should be emphasized that the radiomic feature extraction process was performed following the standards set forth by the Image Biomarker Standardisation Initiative (IBSI) [38].

In contrast to conventional approaches that extract radiomic features exclusively from grayscale representations, our study explores radiomic analysis on both grayscale and RGB image channels to leverage complementary texture and color information. A total of 102 radiomic features per channel were extracted exclusively from the original images only, i.e. no filters were applied. Selected feature classes included: First Order statistics, Shape 2D, GLCM (Gray Level Co-occurrence Matrix), GLDM (Gray Level Dependence Matrix), GLRLM (Gray Level Run Length Matrix), GLSZM (Gray Level Size Zone Matrix), and NGTDM (Neighbouring Gray Tone Difference Matrix). In total, 408 radiomic features were extracted for each lesion.

Thereafter, two deep models were trained and evaluated, including custom DNN (CDNN) and tabular DNN (TDNN) employing TabNet architecture [39]. TabNet is described by its developers as specifically designed to process raw tabular data. It incorporates a feature selection and attention mechanism by design to automatically identify important features while offering local and global explainability. This integrated approach offers the convenience of eliminating the need for supporting methods, such as Lasso [40] for feature selection and SHAP [41] for explanation, and thus streamlining the model development process. Furthermore, TabNet model processes the features in a hierarchical manner through multiple decision steps, where a new subset of features is selected at each step, and partial decisions are made. The model aggregates these partial decisions to produce the final classification. This hierarchical approach allows the model to capture high-level patterns (e.g. overall lesion shape) and finer-grained details (e.g. texture variations). Consequently, the model can adapt to the intricacies of the data which makes it well suited to the heterogeneous nature of skin lesions.

On the other hand, CDNN is a compact deep neural network, which applies batch normalization to the input features, followed by fully connected hidden layers with ReLU activation functions. A dropout layer is incorporated for regularization, and the final classification is performed using a fully connected layer mapping to the seven lesion classes. The custom RBP used CDNN and a more standardized approach to feature selection. This process included eliminating duplicate features, as certain features exhibited identical values across the RGB channels and the grayscale image. Subsequently, the reduction in dimensionality was further refined by applying 5-fold CV Lasso regression, which incorporates L1 regularization to promote sparsity in the feature set. To maintain consistency with the previous DLP, visual explanations were generated using IG.

3.2.5 Hybrid Pipeline

The HP (Figure 2) integrates the deep learning architecture of the best-performing CNN model from the DLP with radiomic pipelines, employing both the TDNN and the CDNN. In both cases, late feature fusion was performed by the feature vec-

tor concatenation, thus leveraging the complementary strengths of both deep learning and radiomic features. In this pipeline, TabNet provides the significance score for the combined features. In the case of custom architecture, IG simultaneously provides an attribution maps from the image input and provides the importance scores of the radiomic features. In this comprehensive framework, along with the prediction and explanations, a segmentation of lesions is provided.

3.2.6 Performance Evaluation

The performance of the segmentation and classification models was evaluated using a set of metrics: accuracy (1), precision (2), recall (3), F1-score (4), area under the receiver operating characteristic curve (AUC-ROC), Dice coefficient (5), and intersection over union (6). The definitions of these metrics are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (3)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (4)$$

$$\text{Dice score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad (5)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \quad (6)$$

where TP denotes true positives, TN true negatives, FP false positives and FN false negatives.

Class-wise metrics, along with aggregated micro, macro and weighted averages, were reported to provide a comprehensive evaluation of the model's performance. For the ROC AUC, the metrics were aggregated in the one-vs.-rest strategy. The class-wise metrics enable a detailed assessment of the model's accuracy in identifying each specific skin lesion class. The micro-average aggregates the performance across all classes, treating each instance equally, whereas the macro-average is the mean of the metrics across all classes, treating each class equally regardless of its frequency. The weighted

average accounts for the class distribution, providing a balanced measure that reflects the relative importance of each class in the dataset.

The metrics reported are presented in the form of MEAN (SD), computed across five independently trained models obtained via 5-fold CV, each evaluated separately on the fixed ISIC-2018 test set. All metrics are normalized to the range of 0–100 to ensure consistency in the representation. Balanced accuracy was excluded from the tables, as it is identical in value to the weighted recall metric, differing only in terminology. This approach limits redundancy and still allows comparison of results with other studies.

3.2.7 Model Interpretability

In the development of the proposed skin lesion diagnosis support system, a principal objective was to guarantee the system's interpretability for the end users. The system incorporates a segmentation neural network that generates binary lesion masks, thereby facilitating clear delineation of lesions for subsequent analysis by experts.

Furthermore, the system provides predictions accompanied by the IG. This choice ensures that the attribution maps remain sufficiently fine-grained to visually inspect the contribution of artifacts (like hair or ruler marks). In our previous benchmarking, no single XAI method consistently dominated across metrics such as Faithfulness, Sensitivity, Localization, Complexity, and Randomness [42], which measure different aspects of attribution quality. Based on these results and our knowledge of the methods, we selected Integrated Gradients, as it allows us to verify whether the model focuses on artifacts rather than the lesion regions, which is critical.

Moreover, an optimal feature selection mechanism is employed to enhance transparency, which strategically simplifies the analysis of feature importance without overwhelming the end user. The system employs radiomic features, offering a well-established medical solution for providing insights into lesion characteristics. Radiomic feature importance can be assessed using two combined approaches: the inherent attention mechanism within TDNN based RBP and HP, and the IG algorithm in the CDNN based RBP and HP. This comprehen-

sive strategy is designed to improve the system's trustworthiness and users' understanding of clinical decision-making procedures, thereby enhancing the system's usability in real-world clinical settings.

3.2.8 Data Investigation

All three pipelines were analyzed separately with and without a hair removal procedure. Unexpectedly, hair removal led to poorer results. Therefore, it was decided to analyze the data more deeply, particularly in terms of the occurrence of various types of artifacts. In this regard, binary labels were prepared for the entire ISIC-2018 Task 3 training (HAM10000) and test datasets for five categories. Prior to the dataset labeling, a consensus on the definition of artifacts was established and agreed upon. This consensus allowed for a standardized labeling approach that focused on the features easily recognizable, even by non-medical specialists, as hair, ruler marks, interaction fluid, and vignettes. An aggregate category, termed „other”, was designated to include various non-skin background elements (as in-ear or arm images), color markings, unusual vignettes, and other anomalous features that did not fit into the predefined categories (examples are shown in Appendix Figure D.1). This category also included ruler marks with text and other atypical or irregular visual artifacts that could be disruptive.

Subsequently, a statistical analysis was performed to determine whether the occurrence of these factors differed between the semantic classes of skin lesions. The Chi-squared test for independence was used, followed by post-hoc test with Bonferroni correction to account for multiple comparisons. P-value less than .05 was considered to be statistically significant. In addition, the Cramér's V effect size was calculated to assess the strength of the association, with values classified as weak (≤ 0.2), moderate ($0.2 < \text{and} \leq 0.6$), or strong (> 0.6).

Changes in predictions were analyzed for a model trained on data with and without the hair removal algorithm applied. Additionally, prediction entropy was examined using Shannon's entropy [43] as a quantitative indicator to measure the uncertainty in the model predictions.

Further analysis was conducted using pixel attribution methods to extract the attention maps,

which were employed to identify and visualize the regions of interest within the skin lesions. These maps allowed for a more detailed understanding of the model's decision-making process and helped to determine whether the presence of artifacts influenced the model's ability to classify the lesions correctly. The analysis aimed to assess potential bias introduced by these artifacts, as their presence could disproportionately affect the model's performance and lead to skewed predictions, particularly in cases where the artifacts were more prevalent in specific semantic classes of skin lesions.

3.3 Implementation Details

3.3.1 Lesion Segmentation Models

EfficientNet-B7, pre-trained on ImageNet, served as the encoder backbone for the segmentation models. The training was conducted with a batch size of 8, for a maximum of 100 epochs, ADAM as optimizer, an initial learning rate of 1×10^{-3} , first and second moment decay rates of 0.9 and 0.999, respectively, and early stopping with a patience of 10 epochs. Dice loss was employed as the optimization objective to enhance segmentation performance.

3.3.2 Custom Lesion Classification

For RBP and HP, after removing duplicate radiomic features, the features were standardized and a nested CV strategy was employed inside the training fold, utilizing Lasso regularization with a maximum iteration count of 1×10^4 and a tolerance threshold of 1×10^{-3} to further reduce dimensionality. During the training phase, online augmentations were applied to the images for both DLP and HP. Stochastic gradient descent (SGD) with a learning rate of 1×10^{-3} was used for model optimization, and the weight decay was set to 1×10^{-3} to regularize the model. The batch size was set to 32 and a weighted random sampler with class weights inversely proportional to class frequency was implemented for RBP, DLP, and HP to address class imbalances. All pipelines were trained end-to-end for up to 1000 epochs, with early stopping triggered after 50 epochs without improvement in the validation loss. Categorical cross-entropy was used as the objective function.

3.3.3 Tabular Lesion Classification

The models were trained on balanced data for each fold by applying an offline oversampling method. The method used was SMOTETomek [44], where SMOTE [45] generates synthetic samples for under-represented classes, and Tomek links remove noisy or overlapping majority class samples. In HP, a two-stage approach was applied. Deep features were pre-extracted using DLP and combined with radiomic features to represent each patient. This joint feature set underwent offline augmentation with SMOTETomek to balance the dataset prior to training.

The training was performed with a batch size of 4096, a maximum of 10000 epochs, ADAM as the optimizer, first and second moment decay rates of 0.9 and 0.999, respectively, and early stopping after 100 epochs without improvement in validation loss. The learning rate was reduced when the validation loss plateaued to allow more effective convergence during the training. Categorical cross-entropy loss was used as the objective function.

In the tabular RBP and HP pipelines, the optimization of hyperparameters was performed using a three-fold CV. The optimized hyperparameters comprised learning rate and the TabNet architectural configuration such as width of the decision prediction layer, width of the attention embedding, number of decision steps, as well as the sparsity regularization coefficient. The hyperparameter search space was carefully designed based on prior empirical findings and dedicated guidelines presented in the original paper [39]. The number of trials conducted for hyperparameter tuning was limited to one hundred.

3.3.4 Model Interpretability

Model interpretability for custom RBP, DLP and HP pipelines was achieved by applying the IG, the method from the XAI family algorithms. A total of 200 integration steps were used, with a uniform baseline applied to the image data and a zero baseline applied to the radiomic features. The uniform baseline was chosen for the image data because it facilitates the identification of attributions not only to the lesion, but also to hair and ruler marks, which are occasionally present in the images as dark or light pixels. Only the positive-sign attri-

butions were visualized in image attribution maps to highlight the regions contributing most significantly to the model's predictions. With regard to the radiomic features, the zero baseline was selected based on the fact that the features were normalized using training data statistics to have a zero or near-zero means in the test set as a result of differences in data distributions. Based on their absolute values, the top 10 radiomic attributions were selected for visualization. In case of tabular models, the attention scores were used to rank the importance of the features used in the model predictions.

3.3.5 Hardware and Software

The experiments were carried out on a workstation comprising an Intel Core i9-14900K CPU equipped with 32GB of RAM and an NVIDIA RTX4090 GPU with 24GB of memory, running on Ubuntu 22.04 distribution. The software was developed using Python version 3.10. The PyTorch deep learning framework (v2.2.1) was employed for model implementation, with GPU acceleration provided by CUDA (v12.4). Hyperparameter tuning was performed using the Optuna Optimization Framework (v4.0.0). For model interpretability, the Captum library (v0.6.0) was used. The statistical analyses, including Chi-squared tests for independence, were performed using the SciPy (v1.9.1) stats module, and corrections for multiple comparisons were implemented using the Statsmodels (v0.14.0) stats module. Radiomics were extracted using the PyRadiomics library (v3.0.1). Random seed initialization (seed=42) was applied consistently to ensure the reproducibility of experiments.

4 Results

All the benchmark results reported in this study were obtained using the official ISIC-2018 Challenge testing data.

4.1 Segmentation Performance

The segmentation performance of the proposed system was assessed using external ISIC-2018 Task 1 test set, with the results summarized in Table 2. Of the models evaluated, U-Net++ without hair removal yielded slightly better segmentation performance. All segmentation models delivered satis-

factory results, showcasing their ability to reliably produce precise lesion masks. Notably, the application of the hair removal algorithm had only a minimal effect on the segmentation accuracy, indicating that the models are robust against common image artifacts. Overall, these segmentation outcomes confirm the effectiveness of the proposed approach in accurately identifying skin lesions, thereby supporting the subsequent classification tasks.

4.2 Classification Performance

The performance of the proposed approaches was evaluated on the ISIC-2018 Task 3 test set. The results of the evaluation are gathered in Table 3 (for AUC-ROC, see Appendix Table A.1). The results were also benchmarked against other studies reported in the literature in order to demonstrate their effectiveness in skin lesion classification. The results reveal that the proposed models achieve competitive performance across all evaluated metrics, with improvements observed in classification accuracy and balanced accuracy (macro recall) compared to the previous studies. Despite the evident lower performance of the RBP compared to DLP, both in the custom and tabular model, the integration of radiomic and deep learning features through multi-modal data fusion in HP improved all classification metrics. Notably, the custom RBP performed better than the tabular RBP in terms of performance. Although the tabular HP sometimes shows better average performance compared to other lesion types, it has higher variability. This improvement underscores the complementary nature of the radiomic and deep features, with the former providing handcrafted and more interpretable features and the latter capturing complex imaging patterns.

The summarized findings suggest that the application of the hair removal algorithm led to a decrease in performance across all evaluated metrics. Contrary to initial expectations, these results may indicate the possibility of a biasing factor resulting from the presence of particular image artifacts or the degradation of the image quality. Notably, radiomic-based models exhibited the least sensitivity to the application of the hair removal algorithm. This can be attributed to the fact that the radiomic features were extracted solely from the segmented lesion region, minimizing the likelihood of includ-

ing artifacts such as hair or interface fluid. In addition, the region of interest used for radiomic feature extraction typically excludes areas containing ruler marks and vignettes, as these artifacts are generally located outside the lesion boundary. As a result, their influence on the radiomic features is minimal or negligible. These findings led us to manually label the dataset for the presence of hair and other task specific artifacts, such as rulers, interface fluids, and vignettes, and then perform quantitative and qualitative analyses.

The average results of class-wise metrics, specifically precision, recall, F1-score, and AUC-ROC, obtained from the 5-fold CV, are shown in Appendix Tables A.2–A5. In general, the results were considered satisfactory for each class, although it was noted that there was a potential for improvement in the *mel* and *akiec* classes.

4.3 Statistical Analysis

A review of the literature and the analysis of the datasets revealed numerous arguments both for and against hair removal from dermoscopic images, often lacking clear statistical evidence. To address this, a comprehensive analysis of the prevalence of artifacts was undertaken to provide a clearer understanding of their impact and to support future research in this area. The occurrence of the identified artifacts on the images from the ISIC-2018 Task 3 training and test datasets is presented in Table 4 (see Appendix Figure B.1 for visual comparison).

In the case of the training set ($n=10015$), benign cases seem to be more frequently associated with the presence of hair in the images, particularly in *nv* (80.4%), *df* (73.9%) and *vasc* (69.0%), compared to other classes, which exhibit hair in 60% of cases. In contrast, benign cases, such as *nv* (10.9%) or *vasc* (12.7%), were not typically associated with ruler marks in the images, whereas the other semantic classes, including malignant *mel*, *bcc*, and precancerous *akiec* had rulers in about 30% of cases. In general, the interface fluid was the most common in malignant cases such as *mel* (24.3%) and *bcc* (19.6%). The presence of vignettes was also more frequent in images of malignant lesions such as *mel* (22.9%) and *bcc* (17.5%), however, vignettes were also present in benign *df* (23.5%) images. Noteworthy, in the training set, the majority class *nv* ($n=6705$) exhibited the largest fraction of images

Table 2. Segmentation models metrics on ISIC-2018 Task 1 test set: 5-fold CV [Mean (SD)].

MODEL *	IoU	DICE	PRECISION	RECALL	F1-SCORE
U-Net ^h	81.14 (0.47)	88.79 (0.31)	85.10 (0.60)	95.54 (0.11)	88.79 (0.30)
U-Net ^{nh}	80.84 (0.42)	88.60 (0.34)	85.08 (0.42)	95.26 (0.82)	88.60 (0.34)
U-Net++ ^h	81.30 (0.34)	88.90 (0.21)	85.50 (0.44)	95.27 (0.08)	88.90 (0.21)
U-Net++ ^{nh}	81.00 (0.43)	88.66 (0.56)	85.40 (0.27)	95.08 (0.76)	88.66 (0.56)
DeepLabV3 ^h	81.02 (0.37)	88.72 (0.23)	85.00 (0.62)	95.55 (0.31)	88.72 (0.23)
DeepLabV3 ^{nh}	81.02 (0.51)	88.62 (0.54)	85.60 (0.71)	94.78 (0.76)	88.62 (0.54)

* The abbreviations used are: *h* for with hair, *nh* for no hair.

Table 3. Aggregated metrics for classification models: 5-fold CV [Mean (SD)].

MODEL *	MACRO ↑			WEIGHTED ↑		
	PRECISION	RECALL **	F1-SCORE	PRECISION	RECALL **	F1-SCORE
RBP _C ^h	45.84 (1.02)	53.80 (1.74)	47.18 (1.28)	71.37 (0.77)	57.64 (1.31)	61.12 (1.09)
RBP _C ^{nh}	45.55 (0.99)	54.15 (0.85)	47.30 (0.82)	71.41 (0.53)	57.91 (1.42)	61.46 (1.15)
RBP _T ^h	25.76 (0.99)	28.05 (0.51)	26.09 (0.81)	51.59 (0.89)	45.42 (2.64)	47.71 (1.53)
RBP _T ^{nh}	23.73 (1.16)	25.86 (1.14)	23.91 (1.04)	51.03 (1.17)	43.75 (1.62)	46.60 (1.13)
DLP _C ^h	72.69 (1.28)	70.20 (1.18)	70.25 (1.09)	80.82 (0.47)	77.60 (0.67)	78.48 (0.53)
DLP _C ^{nh}	69.12 (2.11)	69.97 (1.28)	69.12 (1.93)	79.20 (0.55)	76.88 (0.80)	77.69 (0.66)
HP _C ^h	74.80 (0.84)	73.93 (2.19)	73.62 (1.38)	82.23 (0.41)	79.76 (0.73)	80.48 (0.63)
HP _C ^{nh}	69.38 (1.03)	70.65 (2.34)	69.60 (1.10)	76.88 (1.08)	79.17 (0.40)	77.66 (0.89)
HP _T ^h	73.49 (0.87)	65.19 (0.56)	68.30 (0.40)	80.42 (0.32)	80.38 (0.38)	80.00 (0.26)
HP _T ^{nh}	69.86 (1.07)	61.62 (0.99)	64.68 (0.59)	78.17 (0.30)	77.90 (0.51)	77.79 (0.51)
[24]	—	72.60	—	—	—	—
[26]	—	—	—	—	79.20	—
[25]	—	62.10	—	—	—	—
[23]	—	86.20	—	—	—	—

* The abbreviations used are: *h* for with hair, *nh* for no hair, *C* for custom classifier, and *T* for tabular classifier.

** The MACRO RECALL is equivalent to BALANCED ACCURACY, and WEIGHTED RECALL is equivalent to MICRO ACCURACY.

— indicates that metric was not reported in the paper.

Table 4. Artifact occurrence and Chi-square test results in train and test datasets [n (%)].

ARTIFACT CATEGORY	OCCURRENCE PER CLASS IN THE TRAIN SET (n=10015)							CHI-SQUARE TEST RESULTS
	<i>nv</i> (n=6705)	<i>mel</i> (n=1113)	<i>bkl</i> (n=1099)	<i>bcc</i> (n=514)	<i>akiec</i> (n=327)	<i>vasc</i> (n=142)	<i>df</i> (n=115)	
HAIR	5389 (80.4)	700 (62.9)	710 (64.6)	313 (60.9)	191 (58.4)	98 (69.0)	85 (73.9)	$\chi^2(6) = 356.19$ $p < .001 V = 0.19$
RULER MARKS	734 (10.9)	299 (26.9)	351 (31.9)	161 (31.3)	130 (39.8)	18 (12.7)	40 (34.8)	$\chi^2(6) = 637.37$ $p < .001 V = 0.25$
INTERFACE FLUID	692 (10.3)	271 (24.3)	197 (17.9)	101 (19.6)	73 (22.3)	21 (14.8)	13 (11.3)	$\chi^2(6) = 225.31$ $p < .001 V = 0.15$
VIGNETTE	432 (6.4)	255 (22.9)	203 (18.5)	90 (17.5)	30 (9.2)	18 (12.7)	27 (23.5)	$\chi^2(6) = 421.77$ $p < .001 V = 0.21$
OTHER	137 (2.0)	68 (6.1)	39 (3.5)	30 (5.8)	12 (3.7)	2 (1.4)	0 (0.0)	—
	OCCURRENCE PER CLASS IN THE TEST SET (n=1511)							
	<i>nv</i> (n=908)	<i>mel</i> (n=171)	<i>bkl</i> (n=217)	<i>bcc</i> (n=93)	<i>akiec</i> (n=43)	<i>vasc</i> (n=35)	<i>df</i> (n=44)	
HAIR	788 (86.8)	120 (70.2)	159 (73.3)	62 (66.7)	27 (62.8)	30 (85.7)	34 (77.3)	$\chi^2(6) = 63.04$ $p < .001 V = 0.20$
RULER MARKS	142 (15.6)	55 (32.2)	86 (39.6)	29 (31.2)	14 (32.6)	9 (25.7)	18 (40.9)	$\chi^2(6) = 82.61$ $p < .001 V = 0.23$
INTERFACE FLUID	94 (10.4)	46 (26.9)	56 (25.8)	32 (34.4)	7 (16.3)	6 (17.1)	9 (20.5)	$\chi^2(6) = 73.97$ $p < .001 V = 0.22$
VIGNETTE	63 (6.9)	32 (18.7)	47 (21.7)	15 (16.1)	2 (4.7)	5 (14.3)	9 (20.5)	$\chi^2(6) = 57.20$ $p < .001 V = 0.19$
OTHER	114 (12.6)	42 (24.6)	60 (27.6)	42 (45.2)	5 (11.6)	13 (37.1)	12 (27.3)	—

with hair, and the lowest fractions of images with rulers, interface fluids, and vignettes.

In the case of the test set ($n=1511$), the presence of hair was also common in images of benign lesions, particularly in *nv* (86.8%), *vasc* (85.7%), and *df* (77.3%), compared to other classes being present in approximately 70% of cases. Ruler marks were least frequent in *nv* (15.6%) and *vasc* (25.7%), while in the other classes, their occurrence ranged from 31.2% in *bcc* to 40.9% in *df*. Interface fluid was common in malignant cases such as *mel* (26.9%), *bcc* (34.4%), as well as benign case *bkl* (25.5%). In contrast, its occurrence in other benign cases ranged from 10.4% in *nv* to 20.5% in *df*. The occurrence of vignettes was lower in benign cases such as *nv* (6.9%), *akiec* (4.7%) and *vasc* (14.3%), while other classes, including malignant *mel* (18.7%) and *bcc* (16.1%), ranged from 16.1% in *bcc* to 21.7% in *bkl*. Similarly, in the test set, the majority class *nv* ($n=908$) showed the highest proportion of images with hair and the lowest proportion of images with all artifacts except vignettes, in favour of *akiec*.

In general, the rankings of lesion semantic classes for the four artifacts show general consistency between the training and the test sets, with similar top-ranking classes for each artifact. However, the exact order often differs, particularly for ruler marks, interface fluid, and vignettes. These variations highlight dataset-specific differences, as the training set and the test set are independent datasets collected under different conditions. In both class-unbalanced datasets, the largest fraction of images are *nv* lesions, which often contain hair but not the other artifacts, and this could introduce potential bias in model training.

Finally, the elevated occurrence of the artifacts in the other category may reflect increased complexity within the test set compared to the training data, or imply a potential domain shift. This can be justified, as the ISIC-2018 Challenge Task 3 test set was designed to evaluate model generalization under more challenging and diverse conditions than the training data. This observation may help to clarify the potential negative impact on model generalization, as the model may have difficulty adapting to unseen artifacts, which might be responsible for performance degradation.

The Chi-square test for independence (Table 4) revealed significant differences in the occurrence of hair, rulers, interface fluids and vignettes across lesion semantic classes, for both the train and test sets. Cramér's V effect sizes for these differences ranged from 0.15 to 0.25 in the training set and from 0.19 to 0.23 in the test set, indicating associations at the borderline between weak and moderate strength. Consequently, post-hoc multiple comparison chi-square tests were carried out. The p-values for pairwise comparisons, adjusted using the Bonferroni correction, are presented in Figure 3. The results indicate that, to some extent, both datasets show comparable patterns of similarities and differences in the occurrence of labeled artifacts. The results also suggest that the *nv* class differs the most from the other classes in terms of the presence or absence of artifacts, particularly in the test set.

4.4 Impact of the Hair Removal

A quantitative analysis of the impact of applying the hair removal algorithm on the models performance was investigated. This analysis focused on evaluating changes in the semantic class predictions of lesions and the entropy of the model's output. Specifically, class migrations were analyzed using the Sankey diagram. The class migrations were defined as changes in the predicted lesion types between the model trained and tested on the data without the application of the hair removal algorithm and that trained and tested on data with the algorithm applied. This analysis was conducted using a single-fold HP model. Class migrations shown in Figure 4 demonstrate the flow of three level labels for the test set: ground truth (gt), model predictions without hair removal (h), and predictions with hair removal (nh). The width of the bands indicates the proportion of samples that move between the categories. It can be seen that all the semantic classes contain instances that show consistent behaviour in both scenarios. For the majority class *nv*, there is a visible exchange of predictions between the *nv*, *mel*, and *bkl* classes after the hair removal algorithm is applied. Benign cases such as *df* and *vasc* seem to be unaffected by the algorithm. On the other hand, *akiec* and *bkl* show more false positives, with the primary target of the prediction shifting away from *bcc*. The diagram illustrates that

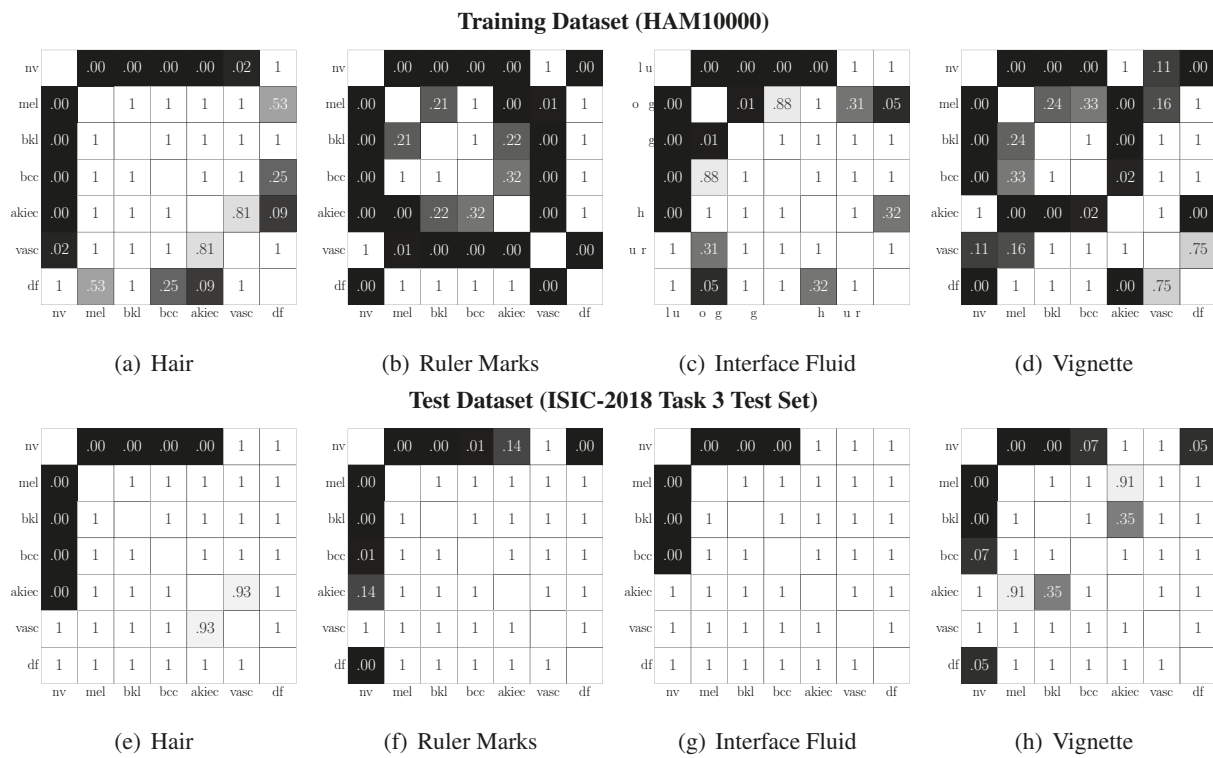


Figure 3. Pairwise comparison Bonferroni corrected p-values for artifact presence in training and test datasets.

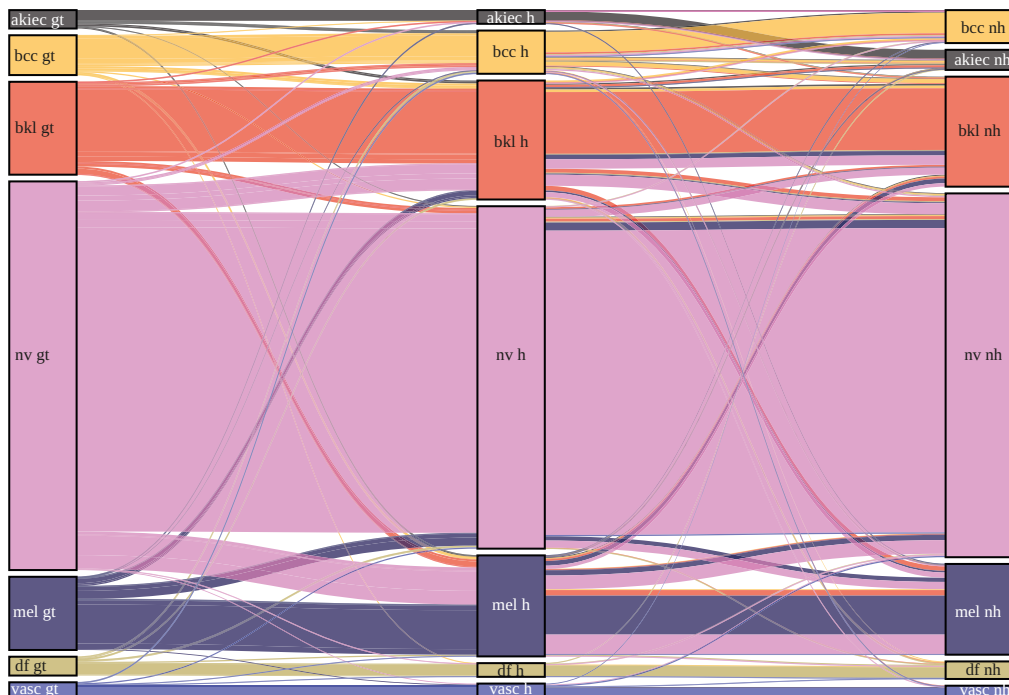


Figure 4. Sankey diagram showing ground truth (gt) labels and test set predictions migration between models trained on data with hair (h) and without hair (nh).

even simple image processing steps can lead to significant and unexpected changes.

The results of further analysis regarding prediction entropy are shown in Figure 6. The box plot presents the impact of the hair removal algorithm on the certainty of the model's predictions across seven skin lesion classes. The hair removal algorithm, originally designed to selectively target hair, has been observed to incidentally impact other thin features, such as ruler marks and lesion boundaries or texture. As a result, in general, the utilization of the algorithm led to an average increase in prediction entropy. In cases where images contained substantial hair coverage, the algorithm unintentionally altered essential diagnostic features. This resulted in a few modified images that, in certain instances, exhibited a higher degree of similarity to lesions from different diagnostic categories. For example, as illustrated in Figure 5, the processed image showed textural patterns similar to the peeling epidermis characteristic of *akiec*, thereby contributing to misclassifications in the algorithm's output. This aligns with previous findings, indicating that the model not only becomes less accurate but also less confident in its predictions following the hair removal.

Notably, classes with a higher prevalence of hair artifacts (see Table 4), such as *nv* and *vasc*, did not exhibit a decrease in classification performance following hair removal, while the remaining classes exhibited a noticeable performance drops (see Appendix, class-wise F1-score Table A.4). This could also be influenced by other artifacts such as ruler marks. Hair removal may falsely clean inputs for classes like *nv*, aiding classification, while removing biased features such as rulers in other classes, potentially degrading performance in models influenced by such artifacts.

4.5 Model Interpretability

A qualitative analysis was performed to assess the interpretability of the model, with a focus on whether the labeled artifacts identified in our study significantly affected the model predictions. Furthermore, the analysis examined the impact of the image feature degradation resulting from the application of the hair removal algorithm on the model's performance. The presence of pixel attributions targeting the hair was occasionally (only for several

test cases, for an example see Figure 7, more in Appendix C) associated with predictions for the *nv* class. It should be noted that hair appeared in a wide range of complex anatomical forms, including different thicknesses, colors, lengths, being shaved or unshaved, and a variety of shapes. These differences have the potential to influence the model's ability to consistently assign the hair to a particular class. Primarily, long and thick hair were among these that received higher attribution scores (Figure 7 and Appendix Figure C.2). In the case of rulers, they generally did not affect the predictions (see Appendix Figures C.3–C.5). However, it was observed that in several cases, the test set contained different types of ruler, including a physical one placed near the lesion to show the scale. In contrast, the training set contained mainly rulers integrated with the magnifying lenses. These physical rulers appeared to contribute significantly to the predictions, probably because the model had not seen them before during the training and misinterpreted them as the skin lesions (see Appendix Figures C.6 and C.7).

On the other hand, in many cases, the application of the hair removal algorithm visibly degraded the texture of the lesion, which resulted in lower prediction probabilities or even altered predictions, with notably reduced pixel attributions observed within the lesion area itself (see Appendix Figures C.3, C.8, C.9 and C.10). Vignettes and interface fluid were not found to contribute much to class prediction. The former, while consisting mainly of black pixels at the image boundaries, did not provide useful information and were among the least common artifacts (see Appendix Figure C.11). Similarly to hair, interface fluid appeared in various forms, ranging from small, homogeneous bubbles to large, heterogeneous gel pools (see Appendix Figures C.4 and C.6). These variations may potentially complicate the model's ability to distinguish meaningful information from misleading patterns introduced by such artifacts.

5 Discussion

All the evaluated segmentation models for the lesion segmentation performed well, both with and without the hair removal algorithm. This may be due to the lesions being typically well-defined, with

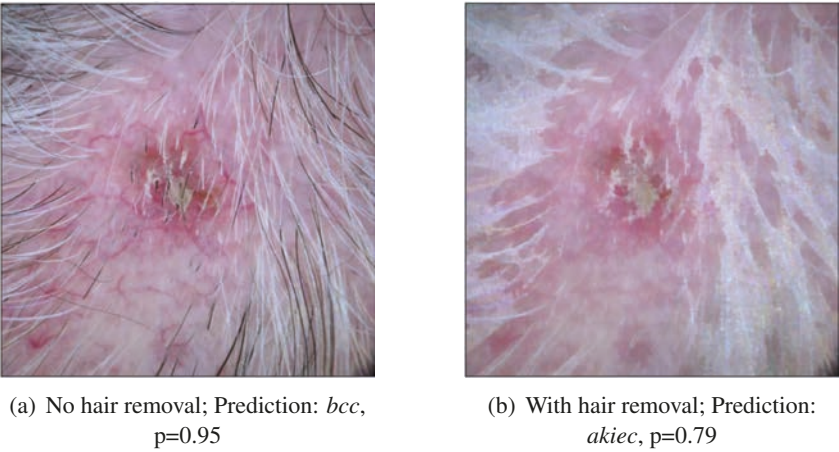


Figure 5. Example of the degrading impact of the hair removal algorithm, illustrating changes in lesion structure and the model’s prediction (ground-truth: *bcc*).

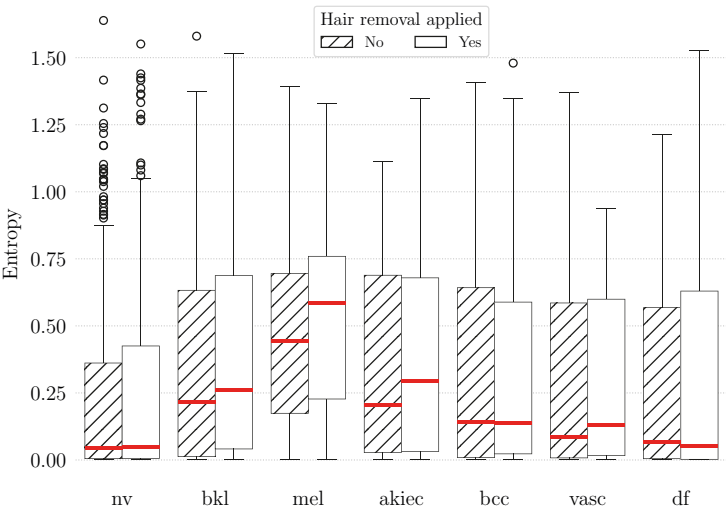
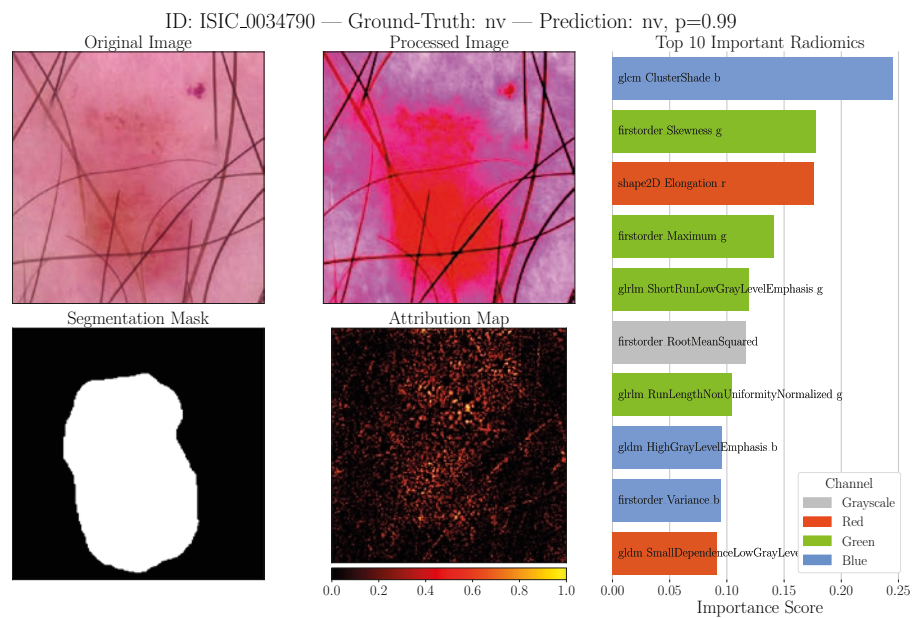
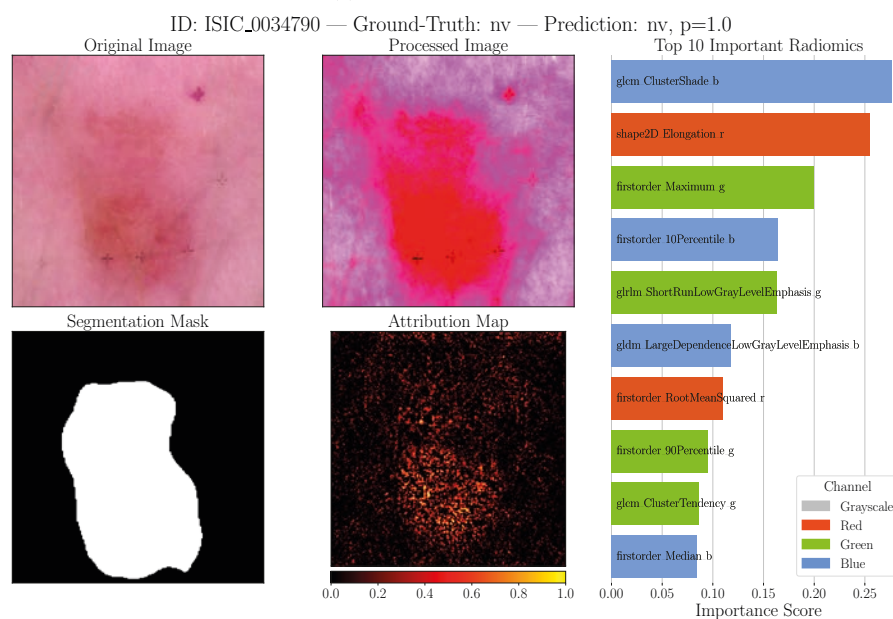


Figure 6. Entropy distribution by predicted class (with and without hair).



(a) Without hair removal



(b) Hair removal applied

Figure 7. SYSTEM INPUT: processed image; SYSTEM OUTPUTS: class probabilities, segmentation mask, attribution map, radiomic importance ranking.

clear boundaries that the models can accurately detect. It was observed that the lesion masks for the models without hair removal were only marginally obscured at the boundaries (see Appendix Figure C.2). Furthermore, the presence of ruler marks, interface fluid or vignettes may not significantly affect the model's ability to segment lesions, as these factors are less likely to come into contact with the lesion area.

Naturally, the RBP showed lower performance due to the limited expressive power of handcrafted features, while the DLP achieved stronger results at the cost of increased opacity. Notably, the radiomics extracted from color channels as proposed in this paper, were more often important for predictions than those derived from the grayscale channel, even though radiomic features were originally designed for gray-level intensity distributions, are therefore most commonly used in that context.

The HP for multi-class classification, combining both deep learning and radiomic image features, demonstrated superior performance compared to its individual counterparts. While this improvement introduces additional complexity, the concerns around interpretability were partially mitigated by employing Integrated Gradients, providing insights into both image-based and radiomic feature contributions, hence helping maintain clinical relevance despite the increased model depth. Moreover, by leveraging lesion segmentation, one can visually compare the attribution maps with the segmentation masks to assess the spatial alignment between model explanations and clinically relevant regions. In contrast to ensembling methods suffering from large number of parameters, which may offer superior performance, our approach may yield slightly lower performance but provides greater interpretability, making it more practical and valuable for medical applications where understanding and validating model predictions is crucial.

The results highlight that the presence of artifacts in medical images can bias model predictions, although this is not always the case. Statistical analysis revealed significant differences in the presence of artifacts in certain classes. However, if artifact appearance or characteristics vary significantly between images, it may be difficult for the model to learn consistent patterns, reducing the likelihood of these artifacts being used to falsely improve the per-

formance. As all images were systematically inspected and labeled for artifacts, the considerable variability in their characteristics became apparent. It is speculated that natural variations in artifact appearance may limit the model's ability to learn reliable patterns, thus reducing the likelihood of artifacts contributing to improved performance.

The algorithms developed to mitigate artifact-related biases, such as hair removal techniques, can reduce the influence of artifacts such as hair. However, these approaches can also have unintended adverse effects, such as the degradation of important image textures, resulting in reduced prediction performance. Overall, the application of hair removal algorithms was found to have a predominantly negative effect on model performance, contrary to the expected outcome of facilitating better lesion analysis. The evidence for this tendency included increased entropy and decreased performance metrics. This highlights the importance of caution when implementing such procedures.

While there is some residual evidence of bias in certain cases, it is not widespread. This suggests that the model might benefit from learning to account for these artifacts.

Moreover, extending this concept beyond hair removal, an alternative, more effective de-biasing strategies could be implemented. For instance, augmenting the training data with synthetic artifacts while preserving lesion texture would help minimize the risk of unintended bias [46, 47]. In the light of the reported results, it is clear that a much more thorough comparative study in this context is needed.

This study has several limitations that should be considered when interpreting the results.

First, the ISIC-2018 dataset does not include annotations for skin tone, Fitzpatrick type, or ethnicity, and prior analyses indicate that the dataset is dominated by images of lighter skin tone [48]. Consequently, the reported performance may not generalize to populations with darker skin tones, where lesion appearance and contrast differ.

Second, the proposed approach relies on a conventional hair removal technique. More sophisticated methods might be utilized. Recent state-of-the-art approaches to hair removal in dermoscopic images emphasize that the overall effec-

tiveness of the preprocessing pipeline is strongly dependent on the accuracy of the hair segmentation stage [49]. Early methods predominantly relied on handcrafted image processing techniques, such as morphological closing or top-hat filtering, derivative-based operators, and conventional edge detectors, followed by interpolation or diffusion-based inpainting to reconstruct occluded regions [50, 51, 52, 53, 54]. While computationally efficient, these techniques are sensitive to variations in hair thickness, color, illumination, and background texture, and segmentation errors at this stage directly propagate to the inpainting step, potentially distorting lesion structures. More recent approaches like Efficient Hair Removal proposed in [49] have shifted toward supervised deep learning-based hair segmentation, leveraging pixel-level annotations to explicitly learn hair appearance from data. However, it requires pixel-level annotations, which are costly and time-consuming to obtain, and such annotations are largely unavailable in existing skin lesion datasets, making their adoption a nontrivial and often impractical requirement in many experimental settings. More sophisticated artifact removal techniques that handle not only hair, but also ruler marks or interface fluid should be considered. Notably, such methods could be extended to remove other artifacts, such as ruler marks or interface fluid.

Another limitation is that recent model families, such as Vision Transformers and hybrid CNN-Transformer architectures, were not explored in this study, and their potential impact on performance and robustness therefore remains unexplored.

Furthermore, the artifact labels were annotated such that each case was labeled by a single annotator, with no overlap between annotators. This does not allow the computation of inter-annotator agreement statistics. While simple artifacts such as hair, ruler marks, interface fluid and vignettes are typically easy to identify, the 'other' artifact category may contain heterogeneous image patterns with greater ambiguity.

6 Conclusions

The hybrid-based pipeline combining deep learning and radiomic image features outperformed individual methods for lesion segmentation and

classification. The system demonstrated strong, state-of-the-art performance proven by the set of dedicated metrics, and robustness, with lesion masks being well-defined and only marginally obscured by artifacts. Additionally, the integration of XAI through integrated gradients enhanced the interpretability of the model, offering insights into the most influential features. Although the identified artifacts can influence the model behaviour, their impact is different, suggesting that the model can adapt to diverse artifact appearances.

The comprehensive set of metrics reported in this paper is expected to facilitate easier and more consistent comparisons of different methods in future research on skin lesion classification and segmentation. Additionally, the novel artifact labels produced for the ISIC-2018 Task 3 dataset are hoped to serve as valuable resources for other researchers.

Future research will focus on incorporating uncertainty quantification into the model and XAI methods to assess model confidence and improve the trustworthiness. In addition, efforts will be directed towards identifying the most robust radiomic features and developing a more refined feature selection process to improve model generalizability and stability. Furthermore, artifact labels will be considered in adaptive preprocessing strategies and artifact-aware training, such as attention mechanisms that can dynamically adapt to artifacts. Finally, the availability of pixel-level artifact annotations could enable a quantitative assessment of whether and to what extent image artifacts introduce bias into model predictions.

Ethics Statement

This study used the publicly available ISIC 2018 dataset, which contains de-identified data collected with ethical approvals (University of Queensland, Protocol No. 2017001223; Medical University of Vienna, Protocol No. 1804/2017). Therefore, no additional approval from the institutional review board was required.

Data Availability Statement

The original ISIC-2018 Challenge dataset is available at challenge.isic-archive.com/data/#2018. The newly labeled metadata can be accessed here² [55], and segmentation masks for the ISIC-2018 Task 3 test set can be accessed here³ [56].

Code Availability Statement

The code for CDNN models, statistical analysis, and decision explanation generation is available at <https://github.com/xkuubix/skin-cancer-classification>. Segmentation and TDNN classifiers definitions and training related code is available at <https://github.com/ChrIstianB/SKIN-LESION-TDNN.git>.

Acknowledgements

The authors also acknowledge the EUCAIM #101100633 consortium for their support. This research was in part supported by the Gdansk Tech IDUB RADIUM grant 14/2/2024/IDUB/III.1a/Ra.

References

- [1] International Agency for Research on Cancer, Cancer TODAY (version 1.1), 2024.
- [2] U. Leiter, T. Eigentler, and C. Garbe, Epidemiology of Skin Cancer, in *Sunlight, Vitamin D and Skin Cancer*, pp. 120–140, New York, NY: Springer, 2014.
- [3] SEER, Melanoma of the skin – Cancer stat facts, 2022.
- [4] F. Nachbar, W. Stolz, T. Merkle, A. B. Cagnetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig, The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions, *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551–559, 1994.
- [5] R. H. Johr, Dermoscopy: alternative melanocytic algorithms—the ABCD rule of dermatoscopy, menzies scoring method, and 7-point checklist, *Clinics in Dermatology*, vol. 20, no. 3, pp. 240–247, 2002.
- [6] J. S. Henning, S. W. Dusza, S. Q. Wang, A. A. Marghoob, H. S. Rabinovitz, D. Polsky, and A. W. Kopf, The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy, *Journal of the American Academy of Dermatology*, vol. 56, no. 1, pp. 45–52, 2007.
- [7] P. Tschandl, C. Rosendahl, and H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Scientific Data*, 2018. Publisher: Nature Publishing Group.
- [8] ViDIR Group, Department of Dermatology, Medical University of Vienna, The HAM10000, 2018. version 4.
- [9] M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks, in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, PMLR, 2017.
- [10] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. Von Kalle, Skin cancer classification using convolutional neural networks: systematic review, *Journal of Medical Internet Research*, vol. 20, no. 10, p. e11936, 2018.
- [11] S. Hagggenmüller, R. C. Maron, A. Hekler, J. S. Utikal, C. Barata, R. L. Barnhill, H. Beltraminelli, C. Berking, B. Betz-Stablein, A. Blum, S. A. Braun, R. Carr, M. Combalia, M.-T. Fernandez-Figueras, G. Ferrara, S. Fraitag, L. E. French, F. F. Gellrich, K. Ghoreschi, M. Goebeler, P. Guitera, H. A. Haenssle, S. Haferkamp, L. Heinzerling, M. V. Heppt, F. J. Hilke, S. Hobelsberger, D. Krah, H. Kutzner, A. Lallas, K. Liopyris, M. Llamas-Velasco, J. Malvehy, F. Meier, C. S. L. Müller, A. A. Navarini, C. Navarrete-Dechent, A. Perasole, G. Poch, S. Podlipnik, L. Requena, V. M. Rotemberg, A. Saggini, O. P. Sanguenza, C. Santonja, D. Schadendorf, B. Schilling, M. Schlaak, J. G. Schlager, M. Sergon, W. Sondermann, H. P. Soyer, H. Starz, W. Stolz, E. Vale, W. Weyers, A. Zink, E. Krieghoff-Henning, J. N. Kather, C. von Kalle, D. B. Lipka, S. Fröhling, A. Hauschild, H. Kittler, and T. J. Brinker, Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts, *European Journal of Cancer*, pp. 202–216, 2021.

²<https://mostwiedzy.pl/en/open-research-data/isic-2018-challenge-task-3-artifactplus,128053118118913-0>

³<https://mostwiedzy.pl/en/open-research-data/isic-2018-task-3-test-set-auto-segmentation-masks,130040415111585-0>

- [12] H. Bhatt, V. Shah, K. Shah, R. Shah, and M. Shah, State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review, *Intelligent Medicine*, vol. 3, no. 03, pp. 180–190, 2023.
- [13] K. Polat and O. Koc, Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all, *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 80–97, 2020.
- [14] S. S. Chaturvedi, J. V. Tembhurne, and T. Diwan, A multi-class skin Cancer classification using deep convolutional neural networks, *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 28477–28498, 2020.
- [15] M. Kumar, M. Alshehri, R. AlGhamdi, P. Sharma, and V. Deep, A de-ann inspired skin cancer detection approach using fuzzy c-means clustering, *Mobile Networks and Applications*, vol. 25, pp. 1319–1329, 2020.
- [16] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chendogen, S. Bhattacharya, and K. Lakshmana, Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, *Scientific Reports*, vol. 12, no. 1, p. 18134, 2022.
- [17] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, Soft attention improves skin cancer classification performance, in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4, (Cham)*, pp. 13–23, Springer, 2021.
- [18] D. N. H. Thanh, V. B. S. Prasath, L. M. Hieu, and N. N. Hien, Melanoma skin cancer detection method based on adaptive principal curvature, colour normalisation and feature extraction with the ABCD rule, *Journal of Digital Imaging*, vol. 33, no. 3, pp. 574–585, 2020.
- [19] Z. Wang, C. Wang, L. Peng, K. Lin, Y. Xue, X. Chen, L. Bao, C. Liu, J. Zhang, and Y. Xie, Radiomic and deep learning analysis of dermoscopic images for skin lesion pattern decoding, *Scientific Reports*, vol. 14, p. 19781, 2024.
- [20] A. C. Foahom Gouabou, J.-L. Damoiseaux, J. Monnier, R. Iguernaissi, A. Moudafi, and D. Merad, Ensemble method of convolutional neural networks with directed acyclic graph using dermoscopic images: melanoma detection application, *Sensors*, vol. 21, no. 12, p. 3999, 2021.
- [21] A. C. Foahom Gouabou, J. Colenne, J. Monnier, R. Iguernaissi, J.-L. Damoiseaux, A. Moudafi, and D. Merad, Computer aided diagnosis of melanoma using deep neural networks and game theory: application on dermoscopic images of skin lesions, *International Journal of Molecular Sciences*, vol. 23, no. 22, p. 13838, 2022.
- [22] A. Bissoto, F. Perez, V. Ribeiro, M. Fornaciali, S. Avila, and E. Valle, Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD titans at isic challenge 2018, 2018. arXiv:1808.08480 [cs].
- [23] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification, *Computer Methods and Programs in Biomedicine*, vol. 193, p. 105475, 2020.
- [24] C. Barata, M. E. Celebi, and J. S. Marques, Explainable skin lesion diagnosis using taxonomies, *Pattern Recognition*, vol. 110, p. 107413, 2021.
- [25] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, Analysis of the ISIC image datasets: Usage, benchmarks and recommendations, *Medical Image Analysis*, vol. 75, p. 102305, 2022.
- [26] C. Barata, V. Rotemberg, N. C. F. Codella, P. Tschandl, C. Rinner, B. N. Akay, Z. Apalla, G. Argenziano, A. Halpern, A. Lallas, C. Longo, J. Malvehy, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, A reinforcement learning model for AI-based decision support in skin cancer, *Nature Medicine*, vol. 29, no. 8, pp. 1941–1946, 2023. Publisher: Nature Publishing Group.
- [27] Anonymous, MSK Dataset, 2018.
- [28] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC), 2019.
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. ISSN: 1063-6919.
- [30] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18, pp. 234–241, Springer International Publishing, 2015.

- [31] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, UNet++: A Nested U-Net Architecture for Medical Image Segmentation, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, (Cham), pp. 3–11, Springer International Publishing, 2018.
- [32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.
- [33] M. Tan and Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019. ISSN: 2640-3498.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- [35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.
- [36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.
- [37] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, and E. Neri, A deep look into radiomics, *La Radiologia Medica*, vol. 126, no. 10, pp. 1296–1311, 2021.
- [38] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. Pfaehler, A. Rahmim, A. U. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, R. J. Spezi, Emiliano and, S. Tanadini-Lang, D. Thorwarth, E. G. Troost, T. Upadhyaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. van Velden, P. Whybra, C. Richter, and S. Löck, The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, *Radiology*, vol. 295, no. 2, pp. 328–338, 2020. Publisher: Radiological Society of North America.
- [39] S. O. Arik and T. Pfister, TabNet: Attentive Interpretable Tabular Learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6679–6687, 2021.
- [40] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [41] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, (Red Hook, NY, USA), pp. 4768–4777, Curran Associates Inc., 2017.
- [42] K. Brzozowski, J. Buler, R. Buler, and M. Grochowski, Comprehensive Quantitative Evaluation of the Performance and Trustworthiness of Deep Learning Models - Skin Lesion Classification Case Study, in *2025 29th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 345–350, Aug. 2025. ISSN: 2835-2807.
- [43] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, no. 3, 1948.
- [44] G. E. Batista, A. L. Bazzan, M. C. Monard, and others, Balancing training data for automated annotation of keywords: a case study., *Wob*, vol. 4, pp. 10–18, 2003.
- [45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, no. 1, p. 321, 2002.
- [46] A. Mikołajczyk, M. Grochowski, and A. Kwasi-groch, Towards Explainable Classifiers Using the Counterfactual Approach - Global Explanations for Discovering Bias in Data, *Journal of Artificial Intelligence and Soft Computing Research*, vol. 11, no. 1, pp. 51–67, 2021.
- [47] A. Mikołajczyk-Bareła, M. Ferlin, and M. Grochowski, Targeted data augmentation for improving model robustness, *International Journal of Applied Mathematics and Computer Science*, vol. 35, no. 1, pp. 143–155, 2025.

- [48] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, Estimating skin tone and effects on classification performance in dermatology datasets, arXiv preprint arXiv:1910.13268, 2019.
- [49] A. Shahzaib, A. B. Siddiqui, N. Anjum, M. U. Rehman, and N. Ramzan, Hair removal and lesion segmentation of dermoscopic images for classification of skin cancer using deep neural networks, *Egyptian Informatics Journal*, vol. 32, p. 100844, Dec. 2025.
- [50] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, Dullrazor®: A software approach to hair removal from images, *Computers in Biology and Medicine*, vol. 27, pp. 533–543, Nov. 1997.
- [51] Q. Abbas, M. E. Celebi, and I. F. García, Hair removal methods: A comparative study for dermoscopy images, *Biomedical Signal Processing and Control*, vol. 6, pp. 395–404, Oct. 2011.
- [52] P. Bibiloni, M. González-Hidalgo, and S. Masanet, Skin Hair Removal in Dermoscopic Images Using Soft Color Morphology, in *Artificial Intelligence in Medicine (A. ten Teije, C. Popow, J. H. Holmes, and L. Sacchi, eds.)*, (Cham), pp. 322–326, Springer International Publishing, 2017.
- [53] A. Huang, S.-Y. Kwan, W.-Y. Chang, M.-Y. Liu, M.-H. Chi, and G.-S. Chen, A robust hair segmentation and removal approach for clinical images of skin lesions, in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3315–3318, July 2013. ISSN: 1558-4615.
- [54] F.-Y. Xie, S.-Y. Qin, Z.-G. Jiang, and R.-S. Meng, PDE-based unsupervised repair of hair-occluded information in dermoscopy images of melanoma, *Computerized Medical Imaging and Graphics*, vol. 33, pp. 275–282, June 2009.
- [55] J. Buler, R. Buler, K. Brzozowski, and M. Grochowski, ISIC-2018 Challenge Task 3 (Artifact-Plus), 2026. [Data set].
- [56] K. Brzozowski, M. Bobowicz, J. Buler, R. Buler, and M. Grochowski, ISIC-2018 Task 3 Test Set (Auto-Segmentation Masks), 2026. [Data set].



Jakub Buler holds a B.Sc. (2022) and M.Sc. (2023) in Control Systems Engineering. Since 2024, he has been working as an Assistant in the Department of Intelligent Control and Decision Support Systems, at the Faculty of Electrical and Control Engineering, Gdańsk University of Technology. He is currently pursuing his Ph.D. (since

2023) at the cutting edge of explainable artificial intelligence and uncertainty quantification. Passionate about innovation, his work focuses on bridging artificial intelligence and medical applications, with a strong commitment to advancing the trustworthiness of AI-based decision support systems to ensure safer and more reliable healthcare solutions.

<https://orcid.org/0009-0008-6684-7446>



Rafał Buler is a Ph.D. student at Gdańsk University of Technology, passionate about advancing the intersection of artificial intelligence and healthcare. His research focuses on developing an advanced framework for integrating diverse data types – images, radiomic features, clinical records, and text – into graph-based models to

enhance predictive accuracy and interpretability in medical AI. With a strong background in machine learning and deep learning, he is particularly focused on establishing and refining strategies for effective graph representation learning, ensuring that the resulting models effectively capture complex patterns and dependencies within multimodal data. Rafał Buler is also an Assistant in the Department of Intelligent Control and De-

cision Support Systems, at Faculty of Electrical and Control Engineering of the Gdańsk University of Technology.
<https://orcid.org/0009-0000-5654-9754>



Krystian Paweł Brzozowski, he is currently a Ph.D. student at Gdańsk University of Technology, specialising in Automation, Electronics, Electrical Engineering and Space Technologies (since October 2025). His research focuses on machine learning for medical applications, with a particular emphasis on generalization theory and privacy-

preserving distributed diagnostic systems. In addition, he is involved in the EUCAIM (European Federation for Cancer Images) project, which is part of the DIGITAL EUROPE programme at the Medical University of Gdańsk. He holds a B.Eng. (2024) and M.Tech. (2025) in Automation, Robotics and Control Systems, both from Gdańsk University of Technology.
<https://orcid.org/0009-0005-6948-6947>



Maria Ferlin received the B.Sc and M.Sc degrees in the field of Control Engineering and Robotics from Warsaw University of Technology (WUT), in 2018 and 2019, respectively. She is currently a PhD student at Gdańsk University of Technology in the field of medical applications utilizing machine-learning approaches. She is also an assistant in the Department of Intelligent Control and Decision

Support Systems, at Faculty of Electrical and Control Engi-

neering of the Gdańsk University of Technology. She leads the project Decision support system for small vessel disease diagnosis, leveraging the synergy of machine learning and radiomics under the PRELUDIUM grant funded by the National Science Centre of Poland. Her research focuses on the applicability of machine learning systems in terms of trustworthiness and interpretability. She also actively participates in non-profit projects aimed at popularising AI-related knowledge and using AI for good.

<https://orcid.org/0000-0001-9286-0670>



Maciej Bobowicz, M.D., Ph.D., serves as an Assistant Professor at the 2nd Department of Radiology at the Medical University of Gdańsk, Poland. He specialises in surgical oncology, focusing on breast cancer diagnostics and treatment. As a researcher and a medical doctor, he advocates the responsible and ethical use of artificial intelligence

in medicine, particularly cancer imaging. His other areas of interest include artificial intelligence in oncology, radiomics, bioinformatics, process automation in radiology, precision medicine in oncology, and emerging cancer diagnostics and treatment technologies. He authored or co-authored over 50 scientific publications, as well as several conference abstracts and papers.

He is involved in four EU-funded projects in the field of artificial intelligence in radiology and oncology and a few international projects with local funding. The largest project, the European Federation for Cancer Images (EUCAIM), is the cornerstone of the European Commission-initiated European Cancer Imaging Initiative, a flagship of Europe's Beating Cancer Plan (EBCP), which aims to foster innovation and deployment of digital technologies in cancer treatment and care to achieve more precise and faster clinical decision making, diagnostics, treatment and predictive medicine for cancer patients.

He is also a member of the clinical working group of the Artificial Intelligence for Health Imaging network, which combines the strengths of 86 European research institutions that are the founders of the EUCAIM project concept. Moreover, he leads clinical teams in two other large-scale European projects 'Eu-CanImage - A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology' and RadioVal - 'International Clinical Validation of Radiomics Artificial Intelligence for Breast Cancer Treatment Planning'. He's also a local PI in the Cinderella Project on 'Clinical Validation of an AI-based approach to improving the shared decision-making process and outcomes in Breast Cancer Patients proposed for Locoregional treatment'. His primary goal is to advance research and medical knowledge for the benefit of his patients by creating trustworthy, fair, and ethical human-centric AI.

<https://orcid.org/0000-0002-3608-1960>



Michał Grochowski, Professor, PhD, graduated from the Faculty of Automatic Control and Robotics at Gdansk University of Technology (GUT). In 2004 he received a Ph.D. degree in Automatic Control and Robotics at the GUT. In 2020 he received a degree in the field of Automation, Electronics, and Electrical Engineering. Currently,

he is a professor and head of the Department of Intelligent Control and Decision Support Systems at the Gdansk University of Technology. He is PI in "European Federation for Cancer Images (EUCAIM)" DIGITAL EUROPE project. His current research is focused on computational intelligence and machine learning methods and their utilization in decision support, data analysis, fault detection, and diagnosis systems, with particular attention to medical diagnostics.

<https://orcid.org/0000-0002-2453-2410>