IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/45AI/2025/4518925/
Vol. 2025 (20), Article ID 4518925

Artificial Intelligence Methods for Multiclass Identification of Skin Diseases*

Maria ROSIAK, Mateusz KAWULOK and Michał MAĆKOWSKI

Department of Distributed Systems and Informatic Devices, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Correspondence should be addressed to: Maria ROSIAK, maria.rosiak@student.polsl.pl

* Presented at the 45th IBIMA International Conference, 25-26 June 2025, Cordoba, Spain

Abstract

Survival in the treatment of skin diseases depends on early detection and accurate diagnosis. There is a need to provide accessible self-assessment solutions for the early recognition of pathological skin lesions. In treatment, early diagnosis is a critical factor for disease remission, effective mitigation of its impact, and improved survival chance. The proposed study focuses on using machine learning methods for analysis and identification of dermatological changes to provide an effective classification system. The dataset used was analysed depending on various factors including age, disease location and gender. In contrast to current solutions based only on two types of changes, the research enables multi-class recognition of pathological skin lesions. Based on the obtained results, it can be stated that the presented solution demonstrated the ability to accurately identify seven skin disease types, achieving an overall accuracy of 80%.

Keywords: image analysis, machine learning, multiclass classification, skin disease diagnosis

Introduction

Conditions manifesting on the skin are the fourth most common cause of all human diseases appearing in Google search results almost 10 billion times annually (Hameed et al. 2019). They affect individuals across all cultural regions and age groups. The highest incidence rate was observed in France, followed by Monaco and then Japan according to data collected from 2011 to 2020 (Li et al. 2020). The standard diagnostic pathway in medicine relies on the visual analysis of lesions and initial clinical examinations, followed by dermoscopic evaluation, biopsy, and histopathological analysis. Dermoscopy is a non-invasive skin imaging technology that enables the visualization of skin structures at the junction of the lower epidermis and the superficial dermis. It is a widely used diagnostic technique that improves the accuracy of diagnosing benign and malignant pigmented skin lesions (82.6% diagnostic accuracy compared to 70.5% without it) (Rosendahl et al. 2011).

Dermatologists usually identify lesions by following the ABCDE (Asymmetry, Border, Color, Diameter, Evolving) rule, which defines the key features of skin lesions as below (Chen et al. 2015):

- The asymmetry of external shape or differential structures within the lesion along at least one axis. This includes contour asymmetry as well as asymmetry in color distribution and dermoscopic structures.

Cite this Article as: Maria ROSIAK, Mateusz KAWULOK and Michał MAĆKOWSKI, Vol. 2025 (20) "Artificial Intelligence Methods for Multiclass Identification of Skin Diseases " Communications of International Proceedings, Vol. 2025 (20), Article ID 4518925, https://doi.org/10.5171/2025.4518925

- The border assessment of the lesion on whether the periphery of the lesion exhibits a sharp, abrupt, gradual indistinct cut-off of the pigment pattern.
- Six colors are considered significant: white, red, light brown, dark brown, blue-grey, and black. Malignant lesions often display more than one shade, whereas benign lesions are typically monochromatic. Color changes include darkening or lightening.
- Differential structural elements or dermoscopic structures includes evaluation of five main features: structureless areas, pigment network, branched streaks (atypical network), dots, and aggregated globules.
- Evolution indicates changes or progression in size, shape, or color.

Using only machine learning it is practically impossible to characterize lesions according to the ABCDE rule due to the lack of public datasets that allow for longitudinal analysis. The training process of algorithms requires a large number of annotated images, but due to data acquisition challenges, high-quality dermoscopic images with reliable diagnoses are limited to only a few disease types, mainly neoplastic diseases. Due to these constraints, researchers have focused mostly on melanocytic lesions while neglecting non-melanocytic pigmented lesions, despite their prevalence in medical practice (Rosendahl et al. 2014; Choi et al. 2019). They are more difficult to classify due to greater clinical heterogeneity, numerous subtypes, variations in severity, and differences in appearance at different stages of progression. There are significant differences within a single category, while similarities exist between categories - similarities in color, texture, scale, and edge contours (Walker et al., 1990).

In recent years one of the biggest barriers in the development of Artificial Intelligence (AI) in medicine is the access to long-term medical and health-related data that represents the diversity of patient populations and disease heterogeneity. This data, although ubiquitous, is not shared and standardized making every resource either too small or too narrow (Barlow 2016).

Environmental conditions surrounding the lesion (e.g. hair, ulceration, veins) also have a significant impact by reducing accuracy in ways that vary depending on the underlying diagnosis. Most modern smartphone cameras are equipped with at least an 8-megapixel sensor with a pixel size of up to 1.5 μm, allowing for high-quality imaging (Codella et al. 2019; Combalia et al. 2022). However, images captured with smartphones exhibit greater variability than dermoscopic images. Deep learning algorithms are highly sensitive to the type of device used for data capture (Goyal et al. 2020). They are typically trained on high-quality datasets, but in the real world, it cannot be assumed that input images will always meet certain standards. Existing networks are susceptible to factors like blurring and noise, making performance degradation not limited to a specific model but common across deep convolutional neural network (DCNN or CNN) architectures (Jaworek-Korjakowska and Kleczek 2018). Smartphone cameras are frequently used under suboptimal conditions by the general population, resulting in variable image quality. Additionally, they lack advanced focal length manipulation capabilities and are not equipped with a white light source for uniform lesion illumination. Images should meet minimum quality requirements, meaning they should be free of shadows or hair covering the lesion and properly focused. Essential factors influencing image quality include the distance between the camera and the lesion, the brightness of the captured image, smartphone characteristics, and the angle of image acquisition (Udrea et al. 2020).

Furthermore, since lesion colors are diagnostic indicators, images should be captured under appropriate lighting conditions. The Color Rendering Index (CRI) for light-emitting diodes (LEDs) sources ranges from 80 to 90, indicating a limited ability to reveal object colors compared to natural light. Comparatively, dermoscopic images are captured with a device consisting of a magnifying lens and a polarized lighting system. The presence of liquid emulsion or cross-polarization filters further reduces skin surface reflections, enhancing visualization of deeper layers (Pan et al. 2008).

Machine learning can be a valuable tool for augmenting human cognitive abilities, considering that the human brain has limited processing capacity while medicine is becoming increasingly complex. Advances in digital pathology - such as increased computational power and cost-effective data storage-facilitate the widespread accessibility of skin lesion detection. The use of AI in this research can provide a more realistic set of diagnostic possibilities regarding the nature of a skin lesion captured in an image.

Related Work

An essential parameter of AI used in diagnostic is reliability and accuracy. Therefore, before a solution can be effectively implemented in real-world conditions, numerous conditions need to be addressed, such as image quality standards, model generalization, the feasibility of algorithm deployment, and the transparency of the decision-making processes. Although AI is commonly used, many aspects - especially the logic behind classification - often remain uninterpretable. This has led to models being frequently referred to as a "black-box" technology (Wang et al. 2019). Relying on treatment decisions made by them contradicts the principles of evidence-based medicine (London 2019).

Nowadays there is a lack of conclusive evidence regarding the safety of using intelligent applications for medical diagnostics and mobile-based AI-driven solutions have not yet demonstrated sufficient accuracy. Systems with embedded algorithms making medical claims require approval. Incorrect recommendations, particularly false reassurances can lead to delays in obtaining a professional medical evaluation (‘DocsRoom - European Commission’ 2025). As of 2018, no automated smartphone applications for melanoma detection had been approved by the U.S. Food and Drug Administration (FDA) (Rat et al. 2018). The CE (Conformité Européenne) marking has been granted to enable the distribution of two AI-powered skin lesion recognition applications in Europe.

The majority of dermatological conditions share similar visual characteristics, making it challenging for individuals without medical training to differentiate them accurately. Consequently, AI may play a crucial role in the early detection of pathological changes due to dermatology heavily relying on morphological features and visual perception, with most diagnoses being based on pattern recognition. Machine learning is able to analyze images efficiency, rapidly and accurately by labeling specific anomalies based on their morphology. In paper (Han et al. 2018), authors applied the ResNet-152 architecture for the classification of images depicting pathological changes. The analyzed cases included squamous cell carcinoma, basal cell carcinoma, actinic keratosis, intraepithelial carcinoma, and malignant melanoma. It was demonstrated that the accuracy of skin lesion recognition may be reduced by factors such as image contrast and the ethnic background of patients. In (Akram et al. 2024), a method was proposed to improve feature representation by integrating multiple deep learning models with an information fusion strategy and theory. Transfer learning was utilized to extract features from Inception-ResNet V2, DenseNet-201, and NasNet Mobile, which were then fused and refined to optimize the feature set. To ensure the preservation of essential and distinctive features while eliminating noise and redundancy, the approach employed an entropy-based binary algorithm.

Previous studies (Combalia et al. 2022) on AI applications in dermatology demonstrated that machine learning can achieve a higher diagnostic classification accuracy than experienced dermatologists. However, these studies did not adequately assess clinically realistic scenarios, such as the model behavior when presented with disease categories not included during training, imaging artifacts or distortions, and images sampled from statistical distributions significantly shifted from the training distributions. An image of a disease not represented in the training data is not assigned to a separate category but is instead misclassified as one of the conditions the model was trained to identify (Combalia et al. 2022).

The primary objective of this study is to assess the performance and reliability of convolutional neural networks in real-world dermatological diagnostics by analyzing their ability to classify skin lesions under clinically relevant conditions. For this reason, this project leverages the potential of CNNs which specialize in visual pattern recognition to propose an automated skin lesion classification solution. Specifically, the research aims to investigate how variations in image source, the presence of out-of-distribution diagnoses, and imaging artifacts influence classification accuracy in the case of skin lesions.

This study seeks to answer the following research questions: To what extent do image artifacts and distortions compromise the accuracy of AI-driven skin lesion classification? What are the limitations of current CNN-based models when encountering out-of-distribution cases, and how does this impact their diagnostic reliability? By addressing these questions, this research aims to provide a comprehensive evaluation of the feasibility, and effectiveness of AI-powered dermatological classification systems, ultimately contributing to the development of more robust and clinically reliable solutions for early detection, prevention, and health promotion in dermatology.

Materials and Methods

The research process involved an analysis of key aspect such as information transfer, rapid accessibility and model accuracy. The proposed solution facilitates communication between a mobile application, a REST API server, and a deep learning model. The system's scope is structured into three fundamental layers:

- presentation layer,
- computational layer,
- data processing layer.

The layers and their working connections are presented in Figure 1.

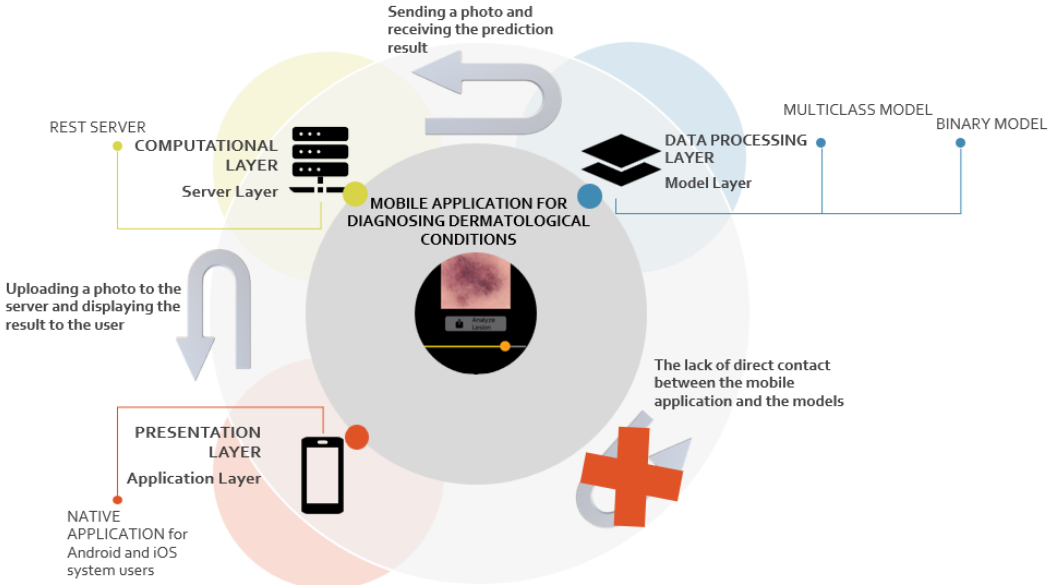

Fig 1. Information flow between the system layers.

Presentation Layer

This layer has been implemented as two native mobile applications with identical functionalities, for iOS and Android phones. They present the results the classification outcome. The research was conducted using an iPhone 8 and 13, as well as a Motorola Moto G6 and Xiaomi Mi 11. The mobile application allowed the following:

- capturing an image,
- guiding the user – instructions were provided on how to properly capture an image of a skin lesion (Figure 2).

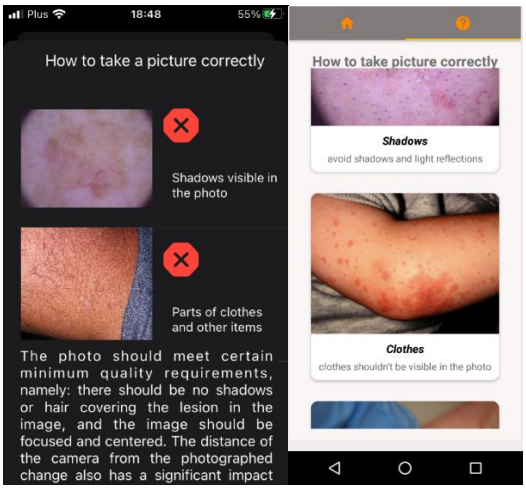

Fig 2. Screenshot showing guidelines for taking a proper photo.

- image modifications – tools for editing were added and included manually adjusting the size of the edited area and applying built-in photographic filters. Additionally, filters that modify contrast, saturation, and the heat scale were available (Figure 3).

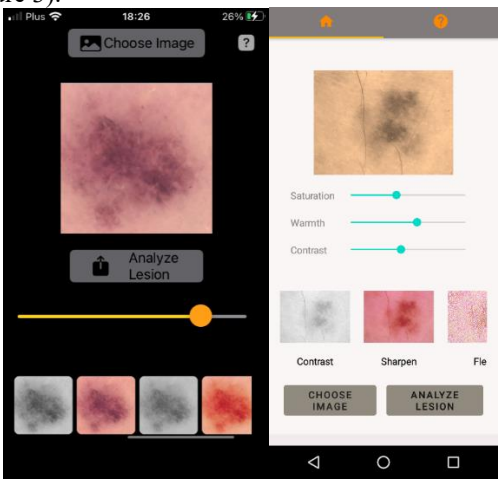

Fig 3. Graphical visualization of photographic filters used in the mobile application.

Computational layer

This layer was responsible for image evaluation hosting and allowed for the analysis of dermoscopic images using resource consuming algorithms. Performing complex calculations on the server helped to minimize network data usage and reduced battery consumption on mobile devices. This was a research assumption because, according to the literature, users tend to cite technical issues as a reason for rejecting applications ('Report: 91 percent would use abandoned apps again if disliked issues addressed' 2025). When an image was taken, it was sent to the server and classified by the model. After processing the data on the server side and obtaining the prediction result, a response containing the outcome was sent back to the mobile application. The server provided feedback on result accuracy while incorporating any updates or modifications to the model directly.

Data processing layer

The skin image, captured for disease identification and classification, underwent pre-processing before being fed as input into the model. The model extracted image features, correlating them with previously trained data to generate a prediction. The input image was decoded, assigned three RGB channels, converted to *float32* within the range [0,1], and resized to the required dimensions. Then the image of the skin lesions was analyzed and evaluated. The multi-

class image classification system leveraged an implemented CNN architecture and used it identify and categorize seven types of skin diseases. The probabilities of each individual type were approximated in order to determine the result class.

Exploratory Data Analysis (EDA)

In this research the ISIC 2018 dataset was used. It was published as a publicly available set under the CC-BY-NC license, as part of the International Skin Imaging Collaboration (ISIC) Grand Challenge Datasets 2018 (Codella et al. 2018). The dataset comprised 10015 dermoscopic images and included a representative set of seven benign and malignant diagnostic categories. The ground truth for labeling was established based on direct observation, expert consensus, or confirmation using in vivo confocal microscopy. All images were categorized according to the classification provided by ISIC.

The ISIC 2018 dataset is characterized by a considerable class imbalance (Figure 4), with a significant predominance of images representing melanocytic nevi (NV) and a very limited number of images depicting dermatofibroma (DF) and vascular lesions (VASC). The remaining classes are more balanced, comprising benign keratosis-like lesions (BKL), melanoma (MEL), basal cell carcinoma (BCC), and actinic keratoses (AKIEC).

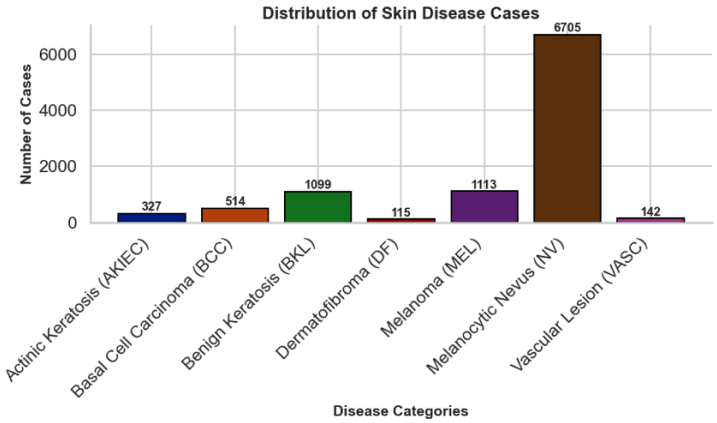

Fig 4. Distribution of Skin Disease Cases in the ISIC 2018 Dataset.

In dermatology, there is no universal standard for image acquisition, processing, transmission, and storage. Therefore, in the research it was important to consider multiple factors, including metadata associated with a given image. Utilizing the *.csv* file provided with the ISIC 2018 dataset the following conclusions have been made:

- Age – The presented age group (Figure 5) has a relatively wide range (approximately 0-85 years), encompassing both children and elderly individuals. The largest group consists of individuals aged 40-60 years, while the least numerous includes children and pre-working-age individuals.

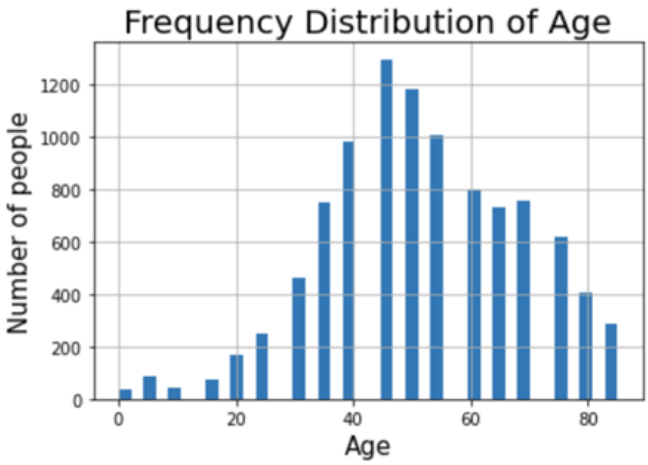

Fig 5. Age distribution of individuals from the ISIC 2018 dataset.

- Body regions most susceptible to specific skin diseases (Figure 6) – Benign keratosis was most commonly diagnosed on the hands, back, and lower limbs, while melanocytic nevi were predominantly located on the lower limbs, back, and trunk. Dermatofibroma was identified on both the lower and upper limbs. Melanoma was most frequently located on the back, upper and lower limbs; vascular lesions on the trunk, abdomen, lower limbs and back; basal cell carcinoma on the back, face and lower limbs. Actinic keratosis was most frequently diagnosed on the hands, upper limbs, and lower limbs.
- Gender and body region – For men, skin lesions were significantly more likely to be located on the back compared to women, where a higher frequency of lesions on the lower limbs was observed. For both genders the most common locations for skin lesions was the same. Less frequently, lesions were found on the trunk, upper limbs, and abdomen. The rarest locations included the scalp, ears, genital area, hands, and peripheral body parts.

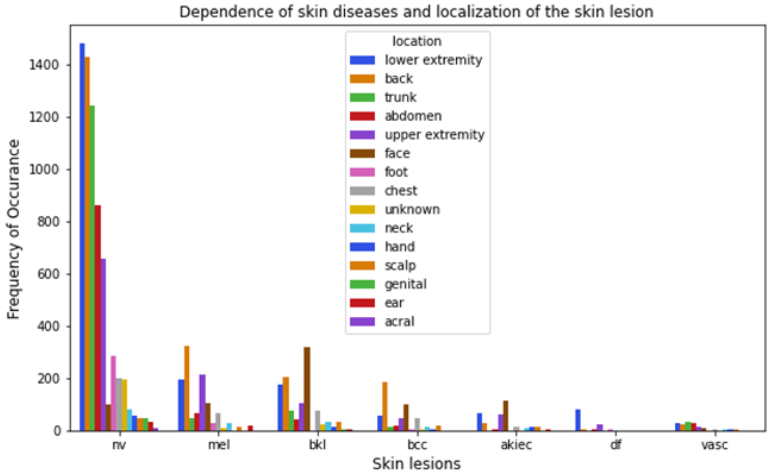

Fig 6. The relationship between the disease entity and the location of the lesion.

- Detection – diagnoses for women were made significantly earlier than for men (Figure 7). For women, disease detection shows a declining trend after the age of 45, whereas for men, the opposite pattern is observed – diagnoses remain at a high level until approximately 75 years of age. The highest incidence rate for both genders is observed in the 45–50 age range. No strong correlations between specific diseases in the ISIC 2018 dataset and gender were identified. For most disease classes, a slight predominance was observed among men. However, it cannot be conclusively determined whether this is due to biological gender differences or simply a result of a higher proportion of men in certain age groups.

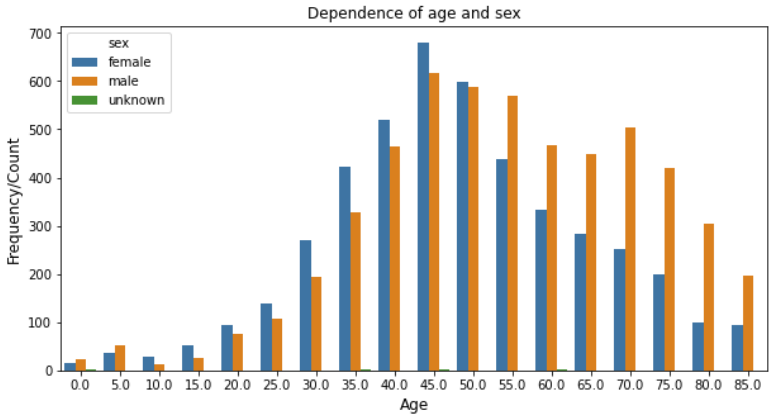

Fig 7. Age and gender dependencies.

Data preprocessing and model parameters

The dataset was randomly partitioned into training and validation sets. The proportion of the data allocated to the validating was 20%, while the remaining images served as the training data. The generator selected images based on a pseudorandom number. The input image size was set to 92×92×3, whereas the original size dermatoscopic images was 450×600 with resizing aiming to reduce memory consumption as well as improve latency. Pixel values were normalized, and the dataset was standardized to obtain a range factor of 0-1. The number of training examples utilized in a single iteration was set to 64. Given that the typical batch size between 32 and 256, various values within this range were tested, and the final value was selected based on performance evaluation and computational efficiency. A significant class imbalance and the small size of some classes presented challenges for training purposes; therefore, data augmentation was employed to mitigate overfitting. To address this issue additional images were generated through rotation and transformation of examples existing in the training set.

Class weights were calculated and assigned based on class sizes, where 0 represented the most frequent class. For instance, there was one occurrence of weight 4 (the most frequent) for every six occurrences of class 6 (the least frequent). Consequently, in the loss function, higher weight values were assigned to less frequent classes, making it a weighted average.

In the proposed architecture, all convolutional layers consisted of the Rectified Linear Unit (ReLU) activation function. The Softmax activation function was used in the output layer, serving as a generalization of the logistic function, ensuring that the predicted probabilities summed to 1. The most important parameters included the following:

- Optimizer: Adam
- Learning rate: the lower bound was set to 0.00001, ensuring relatively slow learning to prevent rapid convergence to a suboptimal solution. Consequently, weight updates per iteration were minimal, necessitating a greater number of epochs.
- Loss function: Categorical Crossentropy, which ensures that the model learns to determine a high probability to the correct class and low probabilities to all others.
- Metrics: Accuracy was used to evaluate model performance.
- Epochs: the model was trained for 100 epochs, with training termination based on the error rate, validation loss, and training loss.
- Callback functions were implemented to trigger specific actions at various training stages to adjust the training loop:
    - EarlyStopping: the loss was monitored at the end of each epoch. If the absolute change in loss was less than 0.02 over 20 epochs it was considered as no improvement and the model was stopped.
    - ModelCheckpoint: monitored the total loss of the validation set, with epoch set predefined intervals.

- ReduceLROnPlateau: the number of weight updates during training was modified by a factor of 0.5 (*lr * factor*) when the metric did not show improvement over 3 consecutive epochs. The metric monitored the loss on the validation set.

Results

After the training phase was completed, the model was deployed to predict the class of images in the test dataset, which had not been used during training. Since ground truth labels were available, these predictions were used to compute performance metrics to assess how well the DCNN classified disease entities. The model was evaluated by monitoring the effectiveness through a dedicated performance assessment module. This module tracked binary cross-entropy loss fluctuations and accuracy trends over successive epochs, alongside additional evaluation methodologies to ensure robust performance analysis.

The dataset used in this phase comprised images collected from smartphone devices sourced from the Dermatological and Surgical Assistance Program (PAD) at the Federal University of Espírito Santo (UFES) – 2298 total samples: 730 - AKIEC, 845 - BCC, 52 - MEL, 244 - NV, and other disease entities (Pacheco, Lima, Salomão, Krohling, Biral, de Angelo, Alves, et al. 2020). In most cases, disease classification was confirmed via biopsy, while the remaining cases were determined through consensus among a team of dermatologists. Integrated testing was conducted to evaluate all available system components working together, simulating real-world conditions.

Model Performance

The confusion matrix (Figure 8) obtained during training allowed to assess how often the model made incorrect predictions. This not only gave insight into the mistakes made but also revealed their types. Moreover, the confusion matrix highlighted the most frequent classification errors covering the three most numerous classes. NV was confused with MEL in 66 cases and with BKL in 51 cases, BKL in 42 cases with NV and 25 with melanoma, while melanoma in 66 cases was recognized as NV and 25 as BKL. An important note were the performance metrics obtained for the least represented classes, where precision exceeded 70% in both cases. The proposed model exhibits a strong diagonal structure in the confusion matrix, indicating that the majority of predictions are correct. The best performance metrics were achieved for the largest class - NV, whereas the poorest metrics were observed for MEL, the second most prevalent class.

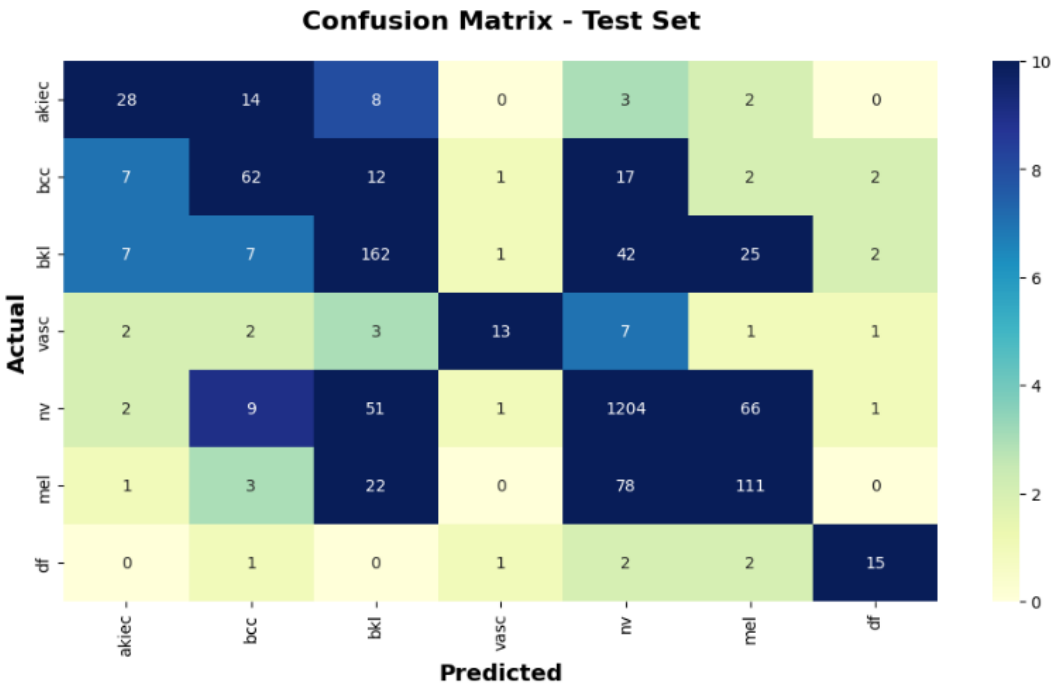

Fig 8. Confusion matrix of the multiclass model.

Additionally, the described training parameters are presented in Table 1.

Table 1: Obtained training parameters.

Class	Precision	Recall	F1-Score	Support
AKIEC	0.60	0.51	0.55	55
BCC	0.63	0.60	0.62	103
BKL	0.63	0.66	0.64	246
VASC	0.76	0.45	0.57	29
NV	**0.89**	**0.90**	**0.90**	**1334**
MEL	0.53	0.52	0.52	215
DF	0.71	0.71	0.71	21
Overall Accuracy	-	**0.80**	-	**2003**
Macro Average	0.68	0.62	0.64	2003
Weighted Average	0.79	0.80	0.79	2003

The precision metrics for individual classes are presented in Figure 9, with results shown before applying augmentation and class weighting (left) and at the final testing phase (right).

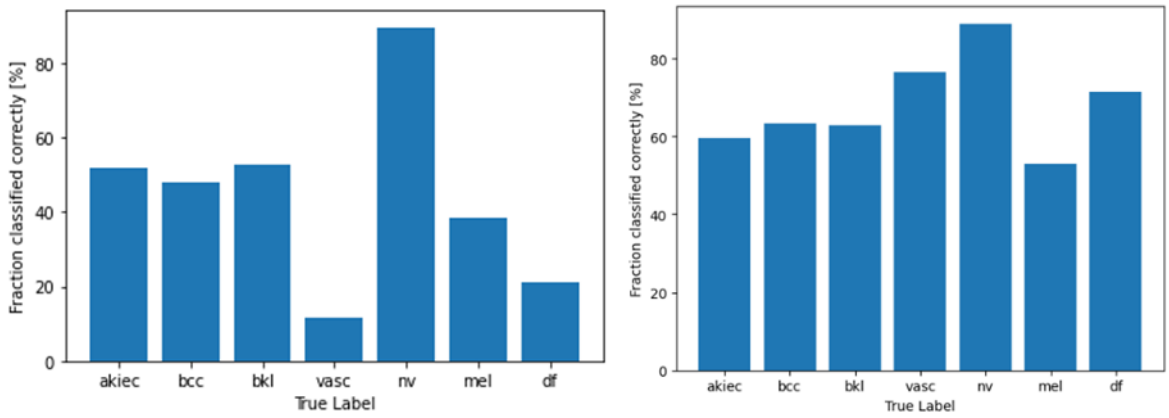

Fig 9. Fraction of correct classifications, before data set augmentation (on the left) and after augmentation (on the right).

Model output results

During the testing stage of the multiclass model, several cases were identified in which the data was misclassified:

- a skin lesion type that the model was not trained to recognize,
- a non-pathological skin condition, such as an irregular tan pattern,
- an image of healthy skin, free from any lesions,
- a photograph that does not depict human skin.

All the above cases were classified as one of the seven categories the model was trained to recognize, regardless of their content. This led to supplementing the system with a binary classifier based on CNN. In this model, one class consisted of images from the ISIC 2018 dataset, while the second comprised of randomly sourced images representing the aforementioned misclassified cases, gathered from multiple datasets available on the Kaggle platform. The binary model's accuracy reached 95.96%, indicating high model effectiveness. The precision was 96.08%, while the recall was 95.82%, and the F1-score was 95.95%. A primary focus was on minimizing Type II errors - cases where the binary model failed to detect one of the seven disease categories - which occurred in 43 instances. If the binary model returned a probability below 0.5, the image was forwarded to the multiclass model. The obtained results are presented in Table 2.

Table 2: Results obtained for the binary classifier.

Metric	Value (%)
Accuracy (ACC)	95.96
Precision	96.08
Recall	95.82
F1-Score	95.95
False Negatives (Type II Errors)	43 cases
High Confidence Predictions	>99.8%
BKL Prediction Probability	≤73%
BCC Correct Predictions	4 cases (>98%)
Melanoma as Second Choice	4 out of 6 cases

Examples of correct predictions for BC are presented in Figure 10. The model correctly identified BCC in four cases with a probability exceeding 98%. Notably, the model achieved accurate predictions despite the presence of artifacts. BKL exhibited significant similarities to other classes, including MEL, NV, and AKIEC. As a result, the

prediction probability did not exceed 73%. In four out of six images, the model identified melanoma as the second most probable class, suggesting notable similarities between these categories.

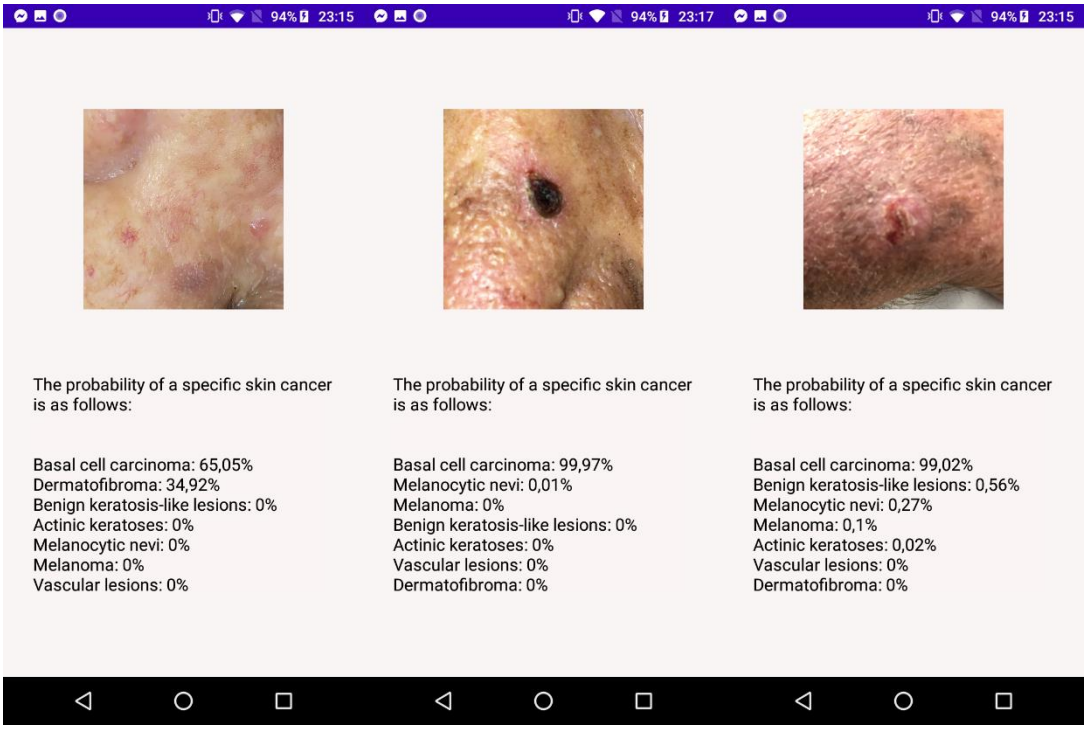

Fig 10. Correct prediction for basal cell carcinoma (BCC).

A major limitation during the second phase research was the model's sensitivity to image quality. A significant challenge arose for images that, in addition to the skin lesion, contained excess background space. Another obstacle to accurate binary model predictions were skin lesions located on highly irregular anatomical surfaces, such as the nose, ears, lips, and areas around the eyes (Figure 11). In such cases, the model frequently made incorrect predictions.

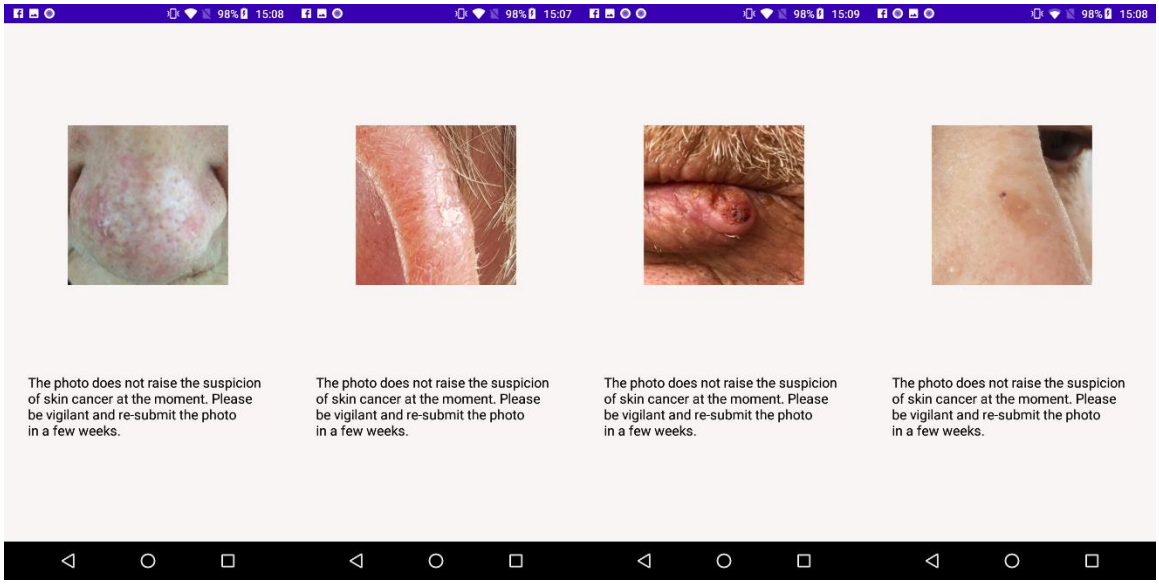

Fig 11. Incorrect prediction at the binary model level.

Model Interpretation Results

Recognition applications are generally based on fractal analysis (Landini 2011). In the context of skin lesions, this pertains to irregularities in the lesion's physical characteristics, such as repeating patterns. Given the increasing emphasis on interpretable models in medicine, understanding the rationale behind model predictions has become a key objective. It has become a legal requirement for mobile applications used for skin lesion recognition to maintain an appropriate level of transparency ('White Paper on Artificial Intelligence: a European approach to excellence and trust - European Commission' 2025). Many high-performing models achieve superior accuracy by learning idiosyncrasies in the data, yet they are often prematurely deployed in real-world use cases. Black-box models remain a major challenge that must be addressed before clinical implementation (Adadi and Berrada 2018; Miotto et al. 2018). ML interpretability bridges the gap between the explicit objectives that models optimize for and the implicit, harder-to-define desiderata in medical decision-making. CNNs composed of distinct hierarchical layers building conceptual blocks. Each layer typically learns a specific feature representation, with lower layers detecting simple features and higher layers identifying complex structures. Figure 12 presents the Local Interpretable Model-agnostic Explanations (LIME) framework, applied in this study for model interpretability. In image-based explanations, the input image is algorithmically segmented into superpixels highlighted in yellow, and the importance of each one of them for classification is determined using linear model. The segmented image is then processed by the LIME algorithm, which generates a classification explanation for the model as a heatmap with blue-to-red color gradients in the range of [-1,1]. Blue areas contribute positively to classification, whereas red areas signify regions that oppose the classification. When compared with medical annotations and literature, the CNN model predictions developed for this project appear to align with human expert knowledge, reinforcing the model's clinical relevance.

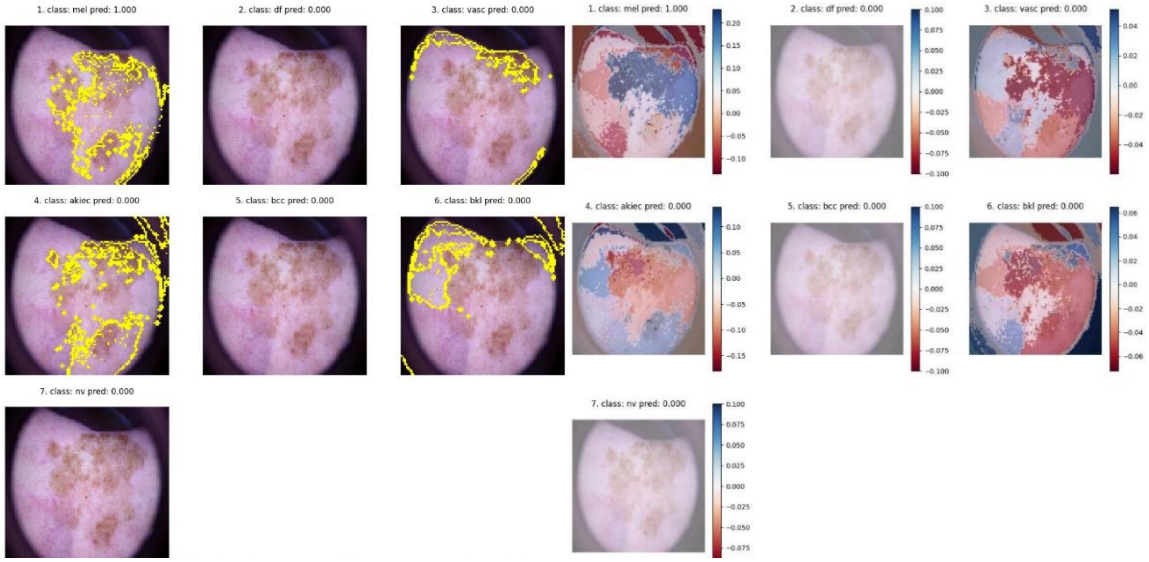

Fig 12. Application of the LIME algorithm to explain the model's predictions.

Conclusion

The comparison of the prediction results and the ground truth on individual examples allowed to identify the classes that were most frequently misclassified. During the research process hypotheses were formulated regarding the causes of such results and the actions to reduce errors were made. Numerous similarities were observed between the BKL, MEL, and NV classes in terms of color, shape, and size. Images representing dermatofibroma were the most unique, having the fewest shared features with other classes. Confusions between BCC, NV and BKL were quite frequent. One important note was the misclassification of melanocytic lesions and malignant melanoma as dermatofibromas when the lesions exhibited a significantly darker color than the majority of cases within these classes.

The creation of monitoring panels for ML models and the implementation of the LIME algorithm contributed to the visual explanation of CNN predictions and a better understanding of how the network makes predictions and why classification errors occur. Formulating hypotheses regarding which features of skin lesions might be important for the model led to a more precise selection of augmentation methods. Prediction on images obtained from mobile devices confirmed observations made during validation on the ISIC 2018 test set. An unusual result was the difficulty encountered with medium-sized classes, rather than the least numerous as initially expected. Dermatofibroma, despite being represented by only 115 images, had, after the most numerous melanocytic nevi class, the best prediction performance.

While CNN-based image analysis has an advantage over human observation in terms of objective and quantitative feature extraction, an obvious drawback is that, unlike human experts, CNNs struggle to distinguish biologically significant features from irrelevant and artifacts ones. A model with a data augmentation function had a greater chance of learning a larger number of discriminative features. Nevertheless, it would be more desirable to include additional data rather than artificially expanding the dataset. Unfortunately, in most medical fields, including dermatology, a large number of manually annotated training images is still not available for a model to learn all texture-based discriminative descriptors for high classification accuracy. Preparing such a dataset requires time, and proper annotation of skin lesions in images demands specialized knowledge. The ISIC 2018 dataset used in the project is relatively new and small compared to other commonly used datasets in computer vision challenges from other fields, such as ImageNet (Deng et al. 2009) and MS COCO (Lin et al. 2014). ISIC 2018 was highly imbalanced across classes, which impacted the final performance of the model.

Presumably, the spatial size of the skin lesion, as well as its color, could also be important factors influencing predictions. Including these features could help reduce the overrepresentation of melanocytic nevi, which was observed. Overall, having a better image database could aid in developing a more reliable algorithm that generalizes skin lesion images more effectively. However, based on the obtained results, it can be stated that the presented system demonstrated the ability to accurately identify seven skin disease units. The conducted tests illustrated the performance of the proposed system and enabled a comparison of the obtained results with those of other studies. The classification of skin disease classes achieved an overall accuracy of 80%, compared to dermatologist detection accuracy, which can range from 75% to 84% on average (Argenziano et al. 2003). In summary, the presented study demonstrates the potential of the proposed system in accurately identifying skin disease classes, achieving competitive performance compared to dermatologists. Despite these promising results, further research will be conducted to refine the model, expand the dataset, and enhance its generalization capabilities for real-world clinical applications.

References

- Adadi, A. and Berrada, M. (2018) ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, 6, 52138–52160.
- Akram, T., Alsuhaibani, A., Khan, M.A., Khan, S.U., Naqvi, S.R., and Bilal, M. (2024) ‘Dermo-Optimizer: Skin Lesion Classification Using Information-Theoretic Deep Feature Fusion and Entropy-Controlled Binary Bat Optimization’, *International Journal of Imaging Systems and Technology*, 34(5).
- Argenziano, G., Soyer, H.P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., De Rosa, G., Ferrara, G., Hofmann-Wellenhof, R., Landthaler, M., Menzies, S.W., Pehamberger, H., Piccolo, D., Rabinovitz, H.S., Schiffner, R., Staibano, S., Stolz, W., Bartenjev, I., Blum, A., Braun, R., Cabo, H., Carli, P., De Giorgi, V., Fleming, M.G., Grichnik, J.M., Grin, C.M., Halpern, A.C., Johr, R., Katz, B., Kenet, R.O., Kittler, H., Kreusch, J., Malvehy, J., Mazzocchetti, G., Oliviero, M., Özdemir, F., Peris, K., Perotti, R., Perusquia, A., Pizzichetta, M.A., Puig, S., Rao, B., Rubegni, P., Saida, T., Scalvenzi, M., Seidenari, S., Stanganelli, I., Tanaka, M., Westerhoff, K., Wolf, I.H., Braun-Falco, O., Kerl, H., Nishikawa, T., Wolff, K., and Kopf, A.W. (2003) ‘Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the internet’, *Journal of the American Academy of Dermatology*, 48(5), 679–693.
- Barlow, M. (2016) AI and Medicine, 1st ed., O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA.
- Chen, L.L., Dusza, S.W., Jaimes, N., and Marghoob, A.A. (2015) ‘Performance of the first step of the 2-step dermoscopy algorithm’, *JAMA Dermatology*, 151(7), 715–721.

- Choi, H., Kang, H., Chung, K.C., and Park, H. (2019) 'Development and application of a comprehensive machine learning program for predicting molecular biochemical and pharmacological properties', *Physical Chemistry Chemical Physics*, 21(9), 5189–5199.
- Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., and Halpern, A. (2019) 'Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC).
- Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., and Halpern, A. (2018) 'Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)', *Proceedings - International Symposium on Biomedical Imaging*, 2018-April, 168–172.
- Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N.R., Liopyris, K., Marchetti, M.A., Podlipnik, S., Puig, S., Rinner, C., Tschandl, P., Weber, J., Halpern, A., and Malvehy, J. (2022) 'Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge', *The Lancet Digital Health*, 4(5), e330–e339.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009) 'ImageNet: A Large-Scale Hierarchical Image Database', *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 248–255.
- DocsRoom - European Commission [online] (2025) available: https://ec.europa.eu/docsroom/documents/17921.
- Goyal, M., Knackstedt, T., Yan, S., and Hassanpour, S. (2020) 'Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities', *Computers in Biology and Medicine*, 127, 104065.
- Hameed, N., Hameed, F., Shabut, A., Khan, S., Cirestea, S., and Hossain, A. (2019) 'An Intelligent Computer-Aided Scheme for Classifying Multiple Skin Lesions', *Computers 2019, Vol. 8, Page 62*, 8(3), 62.
- Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., and Chang, S.E. (2018) 'Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm', *Journal of Investigative Dermatology*, 138(7), 1529–1538.
- Jaworek-Korjakowska, J. and Kleczek, P. (2018) 'eSkin: Study on the Smartphone Application for Early Detection of Malignant Melanoma', *Wireless Communications and Mobile Computing*, 2018(1), 5767360.
- Landini, G. (2011) 'Fractals in microscopy', *Journal of Microscopy*, 241(1), 1–8.
- Li, L.F., Wang, X., Hu, W.J., Xiong, N.N., Du, Y.X., and Li, B.S. (2020) 'Deep Learning in Skin Disease Image Recognition: A Review', *IEEE Access*, 8, 208264–208280.
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014) 'Microsoft COCO: Common objects in context', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5), 740–755.
- London, A.J. (2019) 'Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability', *Hastings Center Report*, 49(1), 15–21.
- Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J.T. (2018) 'Deep learning for healthcare: review, opportunities and challenges', *Briefings in bioinformatics*, 19(6), 1236–1246.
- Pacheco, A.G.C., Lima, G.R., Salomão, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves, F.C.R., Esgario, J.G.M., Simora, A.C., Castro, P.B.C., Rodrigues, F.B., Frasson, P.H.L., Krohling, R.A., Knidel, H., Santos, M.C.S., do Espírito Santo, R.B., Macedo, T.L.S.G., Canuto, T.R.P., and de Barros, L.F.S. (2020) 'PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones', *Data in Brief*, 32, 106221.
- Pan, Y., Gareau, D.S., Scope, A., Rajadhyaksha, M., Mullani, N.A., and Marghoob, A.A. (2008) 'Polarized and nonpolarized dermoscopy: The explanation for the observed differences', *Archives of Dermatology*, 144(6), 828–829.
- Rat, C., Hild, S., Sérandour, J.R., Gaultier, A., Quereux, G., Dreno, B., and Nguyen, J.M. (2018) 'Use of smartphones for early detection of melanoma: Systematic review', *Journal of Medical Internet Research*, 20(4), e9392.

- Report: 91 Percent Would Use Abandoned Apps Again If Disliked Issues Addressed [online] (2025) available: https://martech.org/report-91-percent-use-abandoned-apps-disliked-issues-addressed
- Rosendahl, C., Cameron, A., Tschandl, P., Bulinska, A., Zalaudek, I., and Kittler, H. (2014) 'Prediction without Pigment: a decision algorithm for non-pigmented skin malignancy', *Dermatology Practical & Conceptual*, 4(1), 59.
- Rosendahl, C., Tschandl, P., Cameron, A., and Kittler, H. (2011) 'Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions', *Journal of the American Academy of Dermatology*, 64(6), 1068–1073.
- Udrea, A., Mitra, G.D., Costea, D., Noels, E.C., Wakkee, M., Siegel, D.M., de Carvalho, T.M., and Nijsten, T.E.C. (2020) 'Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms', *Journal of the European Academy of Dermatology and Venereology : JEADV*, 34(3), 648–655.
- Walker, H., Hall, W., and Hurst, J. (1990) 'Clinical Methods: The History, Physical, and Laboratory Examinations', *Geriatric Psychiatry*, 77–121.
- Wang, F., Kaushal, R., and Khullar, D. (2019) 'Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine?', *Ann Intern Med*, 172(1), 59–61.
- White Paper on Artificial Intelligence: A European Approach to Excellence and Trust - European Commission [online] (2025) available: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.