mgr inż. Natalia Krukar[1]
Politechnika Lubelska

https://doi.org/10.29316/9788368103281_13

# QUALITY INDICATORS AND METHODOLOGY FOR EVALUATING ALGORITHM EFFECTIVENESS IN SKIN LESION DETECTION AND SEGMENTATION

## WSKAŹNIKI JAKOŚCI I METODYKA OCENY SKUTECZNOŚCI ALGORYTMÓW W WYKRYWANIU I SEGMENTACJI ZMIAN SKÓRNYCH

**Abstract**

With the global rise in skin cancer cases, Artificial Intelligence-based Computer-Aided Diagnosis (CAD) systems are becoming essential in modern dermatology. This chapter analyzes the methodology for evaluating machine learning algorithms in two key domains: object detection and semantic segmentation. The study challenges the sufficiency of global Accuracy as a primary metric, highlighting its fundamental flaws in the context of imbalanced medical datasets. Based on a literature review, including works on YOLO and U-Net architectures, the clinical significance of the confusion matrix, IoU, mAP, and the DSC is deconstructed. The aim of this study is to define an evaluation standard that balances the technical precision of algorithms with patient safety requirements.

**Keywords:** artificial intelligence, dermatology, evaluation metrics, segmentation, detection, melanoma

[1] e-mail: natalia.krukar@pollub.edu.pl.

**Streszczenie**

Wraz ze wzrostem liczby przypadków raka skóry na całym świecie, systemy komputerowej diagnostyki wspomaganej (CAD) oparte na sztucznej inteligencji stają się niezbędne we współczesnej dermatologii. Niniejszy rozdział analizuje metodologię oceny algorytmów uczenia maszynowego w dwóch kluczowych obszarach: wykrywaniu obiektów i segmentacji semantycznej. Badanie podważa wystarczalność globalnej dokładności jako podstawowego wskaźnika, podkreślając jej fundamentalne wady w kontekście niezrównoważonych zbiorów danych medycznych. Na podstawie przeglądu literatury, w tym prac dotyczących architektur YOLO i U-Net, zdekonstruowano znaczenie kliniczne macierzy pomyłek, IoU, mAP i DSC. Celem niniejszego badania jest zdefiniowanie standardu oceny, który równoważy precyzję techniczną algorytmów z wymogami bezpieczeństwa pacjentów.

**Słowa kluczowe:** sztuczna inteligencja, dermatologia, wskaźniki oceny, segmentacja, wykrywanie, czerniak

## Introduction

Skin cancers are currently one of the most common groups of oncological diseases globally, creating a significant burden for healthcare systems and society. According to data from the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC), over 1.5 million new skin cancer cases were diagnosed in 2022, including approximately 330,000 cases of melanoma, which accounted for nearly 60,000 deaths. These statistics show an upward trend, especially in fair-skinned populations, correlating with lifestyle changes and exposure to ultraviolet (UV) radiation.

The timing of diagnosis is the key factor determining patient prognosis. Early detection of melanoma, when the lesion is limited to the epidermis (in situ) or has low thickness (AJCC Stage I), is associated with a 5-year survival rate exceeding 99%. At this stage, treatment is usually restricted to simple surgical excision. In contrast, diagnosis at an advanced stage (Stage IV), when distant metastases have occurred, drastically worsens the prognosis. Historical survival rates hovered below 15-20%, although modern immunotherapies (e.g., checkpoint inhibitors) are improving these statistics, albeit at a significant cost (Gevorgyan, 2025).

The economic aspect is an inseparable part of the discussion on diagnostics. The cost of treating advanced melanoma is disproportionately higher than treating early lesions. Studies indicate that treating a Stage IV patient can be over twenty times more expensive than Stage I, translating to billions of dollars annually in the United States alone (Crealey et al., 2023). Therefore, early warning and precise diagnostic systems are not merely clinical tools but strategic elements of health policy.

A shortage of dermatologists, combined with the growing number of patients requiring screening, has created a gap that is increasingly being filled by artificial intelligence algorithms. The evolution from classical image processing methods based on handcrafted feature extraction (such as asymmetry or texture) to Deep Neural Networks (DNN), and specifically Convolutional Neural Networks (CNN), marked a turning point. Modern architecture, such as the You Only Look Once (YOLO) series in detection tasks or U-Net and DeepLab in segmentation tasks, achieve performance comparable to, and sometimes exceeding, that of dermatologists in experimental settings (Elshahawy et al., 2023). These algorithms do not merely classify lesions as benign or malignant but perform two fundamental computer vision tasks:

- Object Detection – Answers the question "Where is the lesion?". This task involves localizing the lesion in an image (often macroscopic or dermoscopic) and marking it with a bounding box. This is crucial in triage systems where the algorithm must identify a suspicious area in Total Body Photography.
- Semantic Segmentation – Answers the question "What is the exact shape of the lesion?". Here, the model assigns a label (e.g., "lesion" or "background") to every pixel in the image. Precise boundary delineation is essential for assessing the dermoscopic structure of the lesion, such as border irregularity, which is one of the main criteria in the ABCD rule (Asymmetry, Border, Color, Diameter).

Despite the enthusiasm surrounding AI implementation, issues of trust and verifiability arise. How do we measure if an algorithm is "good"? The traditional "Accuracy" metric can be misleading in medicine. In a dataset where 95% of cases are benign moles and only 5% are melanomas, a model that ignores all malignancies and classifies everything as "healthy" will achieve

an impressive 95% accuracy while being clinically useless and dangerous (Moffat, 2024).

The aim of this paper is to decompose and analyze advanced evaluation metrics that allow for reliable model verification under conditions of strong class imbalance and high risk of error. Specific metrics for detection (mAP, IoU) and segmentation (Dice, Jaccard) will be analyzed, taking into account the specificity of dermatological data, such as the HAM10000 or ISIC Archive datasets.

## The Problem of Data Imbalance in Dermatology

The foundation of any evaluation of classification and detection models in medicine is the confusion matrix. It is a contingency table that compares model predictions against the Ground Truth, typically defined by histopathological examination (biopsy) or expert consensus. The visual structure of this tool, distinguishing between predicted and actual classes, is presented in Figure 1.

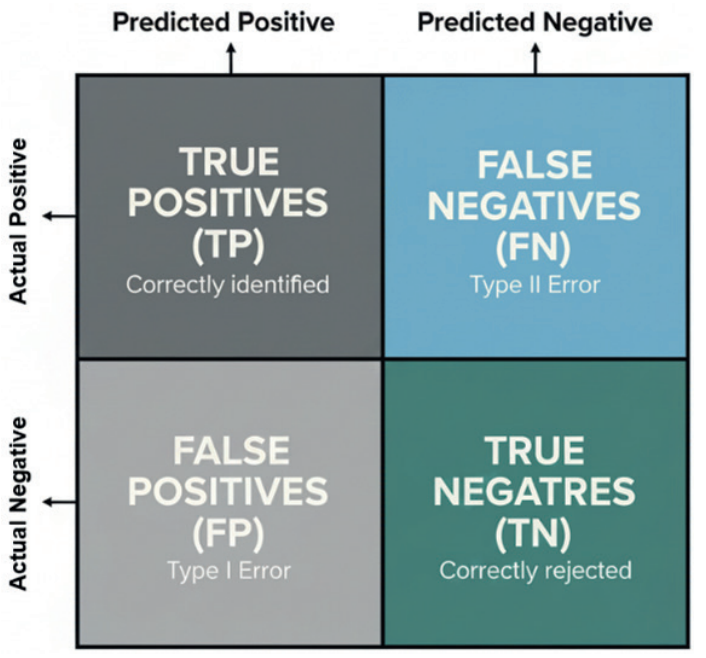



**Figure 1.** Example Confusion Matrix
Source: own study.

In a binary context (malignant vs. benign), we distinguish four states whose medical interpretation is crucial for understanding further metrics (Natha, RajaRajeswari, 2024; Rahi, 2025). In the context of skin lesion diagnostics, these states carry distinct clinical implications:

- True Positive (TP) – The model correctly identifies a malignant lesion (e.g., melanoma) as malignant. This represents a diagnostic success. The patient is referred to for further treatment (excision), which saves lives in cases of early detection. In segmentation tasks, this refers to lesion pixels correctly classified as tumor.
- True Negative (TN) – The model correctly identifies a benign lesion (e.g., a nevus) as benign or background as background. The patient avoids unnecessary stress and invasive procedures. In screening systems, a high TN rate is crucial for cost-effectiveness, preventing healthcare system overload from false alarms.
- False Positive (FP – Type I Error) – The model incorrectly classifies a benign lesion as malignant (false alarm). In detection, this implies marking healthy skin or artifacts (e.g., hair, rulers) with a bounding box. While less critical for survival than FN, this error generates tangible financial and psychological costs. An unnecessary biopsy incurs additional costs, risks of complications, and scarring. On a population scale, a high FP rate leads to resource waste and "alarm fatigue" among physicians.
- False Negative (FN – Type II Error) – The model fails to detect an existing malignant lesion or classifies it as benign. This is a critical error. Missing a melanoma allows for its continued growth and potential metastasis. The cost of such an error is measured not only in currency (treating advanced cancer is vastly more expensive) but primarily in Years of Potential Life Lost (YPLL). Therefore, medical algorithms are optimized primarily to minimize FN (Alexandrescu, 2009).

A concise summary of these states, along with their direct consequences for the patient and the healthcare system, is provided in Table 1.

**Table 1.** Interpretation of the confusion matrix in the context of skin cancer

| State | Medical Interpretation | Consequences for Patient and System |
|---|---|---|
| **TP (True Positive)** | Correct detection of a malignancy. | Early treatment, high cure rate (survival >99% in Stage I). |
| **TN (True Negative)** | Correctly ignoring a benign lesion. | No intervention, patient peace of mind, resource saving. |
| **FP (False Positive)** | False alarm (healthy skin considered cancer). | Unnecessary biopsy (cost, scarring), patient stress, burden on dermatologists. |
| **FN (False Negative)** | Missing a malignancy. | Delayed diagnosis, progression to Stage IV (metastasis), high mortality, drastic increase in treatment costs. |

Source: own study.

## Visualizing the Class Imbalance Problem

In analyzing dermatological data, it is crucial to understand that these four values are not equivalent. In a typical screening dataset, the ratio of benign to malignant lesions can be 10:1 or even 100:1. This leads to the Accuracy Paradox.

The traditional Accuracy metric is defined by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Consider a scenario with 100 patients, where only one has melanoma. A model that diagnoses everyone as "healthy" (TP=0, TN=99, FP=0, FN=1) achieves 99% accuracy. However, its sensitivity (the ability to detect the disease) is 0%. Consequently, in scientific reports and clinical validations of AI systems, Accuracy is treated as a supplementary metric rather than a decisive one (Çetinel, 2024) (Pope et al., 2024).

## Metrics in Object Detection Tasks

In detection tasks, implemented by architectures such as YOLOv5, YOLOv8, or EfficientDet, the goal is not only classification but primarily localization. The model output consists of a Bounding Box, a class label,

and a prediction confidence score. Evaluating the quality of such predictions requires metrics that account for geometry.

**Intersection over Union (IoU)**

Intersection over Union is a fundamental metric quantifying the degree of overlap between the predicted bounding box and the reference box (Ground Truth).

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Where $A$ is the bounding box predicted by the algorithm, and $B$ is the bounding box annotated by the dermatologist. A value of 1.0 indicates perfect overlap, while a value of 0.0 indicates the absence of any shared area. In object detection, IoU serves as a threshold criterion. If the IoU value exceeds a predefined threshold (typically $> 0.5$), the detection is considered a True Positive (TP). Otherwise, it is treated as a false alarm (False Positive - FP), even if the class label is correct (Ultralytics, 2025b) (Shah, 2023).

**Precision and Recall**

After determining which detections constitute True Positives (TP) and False Positives (FP) based on the IoU threshold, two key metrics are calculated.
Precision:

$$Precision = \frac{TP}{TP + FP}$$

This metric indicates the reliability of the model's predictions. High precision means that when the model marks a skin lesion, there is a high probability that it is indeed a lesion, rather than a shadow or a benign nevus.
Recall:

$$Recall = \frac{TP}{TP + FN}$$

This measures the proportion of all existing lesions successfully identified by the model. In dermatology, this is the most critical safety indicator. A model with low Recall "misses" lesions, which is unacceptable in a clinical setting.

**Precision-Recall Curve and mean Average Precision (mAP)**

Since there is a natural trade-off between Precision and Recall (increasing recall by lowering the detection threshold usually increases the number of false alarms, thereby lowering precision), single values do not provide a complete picture. The solution is mAP.

Average Precision (AP) is the area under the Precision-Recall curve (Area Under Curve - AUC PR) for a given class. It is calculated as an integral (or, in practice, a discrete weighted sum):

$$AP = \int_0^1 p(r)\,dr$$

mAP is the arithmetic mean of AP values for all classes (e.g., melanoma, basal cell carcinoma, nevus). Two main standards for reporting mAP are distinguished in scientific literature, as evident in studies on YOLO models (Ultralytics, n.d.) (Ultralytics, 2025a):

- **mAP@50** – A "liberal" metric indicating the correctness of the lesion localization itself (IoU > 0.5). In screening studies, this is often sufficient, as the primary goal is simply to identify the lesion for further analysis.
- **mAP@50-95** – A "strict" metric (COCO standard), calculated as the average over IoU thresholds from 0.5 to 0.95. It rewards not only the detection of the lesion but also its highly precise localization (bounding box accuracy).

**F1-Score in Detection**

As a harmonic mean, the F1-Score is useful for rapid model comparison, providing a single value that evaluates the balance between Precision and Recall:

$$F1\ \ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In the context of detecting small objects (such as early-stage skin lesions), the F1-Score is often more informative than Accuracy because it heavily penalizes low values in either component (e.g., very low recall despite high precision (Khandaker et al., 2025).

**Metrics in Semantic Segmentation Tasks**

Segmentation is a task involving a higher level of granularity. Here, the objective is not a bounding box but a binary mask that distinguishes every pixel of the lesion from healthy skin. Models such as U-Net, DeepLabV3+, or newer transformer-based architectures (e.g., Swin-Unet) are evaluated using set-based metrics.

**Dice Similarity Coefficient (DSC)**

DSC is the most widely used metric in medical image segmentation. It measures the degree of overlap between two sets of pixels: the predicted set $A$ and the ground truth $B$.

$$DSC(A, B) = \frac{2 * |A \cap B|}{|A| + |B|} = \frac{2 * TP}{2 * TP + FP + FN}$$

DSC is normalized (0-1) and intuitive. A value of 1 indicates identical masks. The formula emphasizes the intersection (TP), which is desirable in medical applications. It is the standard in ISIC segmentation challenges. In studies comparing various architectures on hybrid datasets, DSC served as the primary quality indicator. Models such as U-Net++ achieved scores

exceeding 0.90, demonstrating high concordance with expert annotations (Çetinel et al., 2024).

**Jaccard Index**

The Jaccard Index (equivalent to IoU in segmentation contexts) is a stricter metric than the DSC:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

The relationship between these metrics is defined by the equation:

$$DSC = \frac{2J}{1+J}$$

The relationship $J < DSC$ consistently holds (except when both equal 1). This implies that the Jaccard index penalizes errors (FP and FN) more heavily. In research practice, both values are typically reported; however, the Jaccard index is often preferred to avoid over-optimistic results, particularly when segmenting small lesions where every erroneous pixel carries greater weight relative to the total lesion area (Moffat, 2024; Le et al., 2023).

**Limitations of Pixel Accuracy in Segmentation**

Similar to classification tasks, accuracy in segmentation can be calculated at the pixel level (Pixel Accuracy):

$$PixelAccuracy = \frac{TP_{pix} + TN_{pix}}{Total_{pix}}$$

However, this metric is highly misleading in dermatological imaging. A skin lesion often occupies only a small portion of the image (e.g., 5%). A model predicting "background only" (an all-black mask) would achieve a Pixel Accuracy of 95%, while having zero diagnostic value ($Dice = 0$). For this reason, professional scientific literature is moving away from this metric

in favor of the Dice and Jaccard coefficients (Ultralytics, 2025b; Müller et al., 2022).

**Hausdorff Distance (HD) – Clinical Boundary Fidelity**

Overlap-based metrics (such as Dice and IoU) suffer from a significant clinical drawback: they are insensitive to local shape deviations as long as the overall area matches. However, in oncological surgery, where a single millimeter determines the resection margin, the maximum error distance is paramount.

The Hausdorff Distance (HD) metric addresses this issue and is defined as:

$$d_H(A, B) = max\left\{\sup infd(a, b), \sup infd(a, b)\right\}$$

In practice, the 95th percentile of this distance (HD95) is used to eliminate the impact of isolated outlier pixels (noise). If a model achieves a high Dice score (e.g., 0.95) but a high Hausdorff error, it indicates that while the lesion was generally located correctly, there is a significant deviation in a specific area (e.g., the model missed a tumor infiltration). Consequently, in automatic surgical margin planning systems, HD95 takes precedence over the Dice coefficient (Mustafa, Alquran, 2025).

**ROC Curve and Area Under the Curve (AUC)**

In the evaluation of diagnostic systems, single metrics such as sensitivity or precision are dependent on the selected probability threshold. To assess the model's global discriminatory power, independent of the chosen threshold, the Receiver Operating Characteristic (ROC) curve is employed. This curve plots the relationship between Sensitivity (True Positive Rate) and Specificity (expressed as $1 - Specificity$ or False Positive Rate). An ideal classifier is located at the top-left corner of the plot (100% sensitivity, 0% false positives).

### Area Under the Curve (AUC)

This metric ranges from 0.5 (random guessing) to 1.0 (perfect classification). In the context of dermatology:

- $AUC > 0.9$: The model is considered excellent, often outperforming visual assessment by primary care physicians.
- $0.8 < AUC < 0.9$: A good model, clinically useful as a decision support tool.
- $AUC < 0.7$: A model with low diagnostic value.

AUC analysis is a particularly valuable metric for imbalanced datasets, as it remains insensitive to class proportions in the test set, in contrast to Precision or the F1-Score.

### Comparison of Metrics

The analysis of performance metrics presented in the previous sections highlights the complexity of evaluating AI systems in dermatology. No single metric is universally applicable; each possesses specific strengths predisposing it to specific clinical applications (e.g., Recall in screening contexts), but also significant limitations (e.g., Dice's sensitivity to lesion size). To synthesize this knowledge and facilitate the selection of the appropriate criterion depending on the research goal, Table 2 provides a concise comparison of the discussed indicators. This overview includes not only mathematical formulations but primarily the medical interpretation of advantages and disadvantages, which is crucial for accurately assessing the model's clinical utility.

**Table 2.** Metrics Comparison

| Metric | Advantages in Dermatology | Disadvantages / Limitations |
|---|---|---|
| Accuracy | Intuitive; provides a general overview of model effectiveness. | Misleading with imbalanced data. A high score may mask missed cancer detections (False Negatives). |
| Recall (Sensitivity) | Crucial for safety. Measures the ability to detect every cancer case. | Does not account for false alarms (False Positives). On its own, it may encourage the "over-flagging" of lesions. |

| Metric | Advantages in Dermatology | Disadvantages / Limitations |
|---|---|---|
| Precision | Measures the reliability of alerts. Important for reducing the number of unnecessary biopsies. | Ignore missed cases (False Negatives), which poses a significant risk in oncology. |
| F1-Score | Balances Precision and Recall. A robust measure for uneven class distributions. | Treats FP and FN errors equivalently, which is not always desirable in medicine (where the cost of FN > FP). |
| Dice (DSC) | The gold standard in medical competitions (e.g., ISIC challenges). It is intuitive (0-1 range) and emphasizes the correctness of the overlap (True Positives), a priority in pigmented lesion segmentation. | Less strict than Jaccard. Exhibits high sensitivity to very small skin lesions (size bias) – an error of just a few pixels on a small mole drastically lowers the score, which does not always reflect actual clinical utility. Fails to detect the „smoothing" of irregular borders. |
| IoU (Jaccard) | More rigorous than Dice, penalizing False Positive and False Negative errors more heavily. Better reflects the „worst-case" overlap, increasing the safety margin when evaluating algorithms for surgical boundary definition. | Numerical values are always lower than or equal to DSC (), which non-technical stakeholders may misinterpret as „poorer" performance. Like Dice, it is an area-based metric and does not precisely evaluate shape fidelity (border irregularities), requiring supplementation with distance-based metrics (e.g., Hausdorff). |
| mAP | Comprehensive evaluation of detection across various confidence thresholds. | Complex clinical interpretation. Requires analysis at various IoU thresholds (e.g., @50 vs @50-95). |

Source: own study.

## Discussion

The evaluation of AI algorithms does not occur in a vacuum. Numerical results (*mAP, DICE*) are intrinsically linked to the quality and characteristics of the test data. In dermatology, specific challenges must be considered when interpreting these results. Datasets such as HAM10000 or the ISIC Archive reflect clinical reality, where benign lesions (nevi, seborrheic keratosis) occur significantly more frequently than melanomas or skin cancers. In the HAM10000 dataset, nevi constitute the vast majority (over 6700 cases),

while vascular lesions or dermatofibromas number only a few hundred each (Adebiyi et al., 2024; ISIC, n.d.). Training on such data without appropriate techniques (e.g., augmentation, weighted loss functions like Focal Loss) leads to a bias toward the dominant class. The algorithm learns that it is statistically advantageous to predict a "benign lesion." This results in high Accuracy but low Recall for the minority class (melanoma). Therefore, performance analysis must be conducted on a class-wise basis rather than globally (Pope et al., 2024; Alam et al., 2022).

In supervised learning, it is typically assumed that labels (Ground Truth) are indisputable. In dermatology, this assumption is flawed. Studies indicate that inter-rater agreement among dermatologists in assessing dermoscopic features can be moderate (Cohen's Kappa coefficient in the range of 0.5-0.6). There exists a group of lesions (e.g., dysplastic nevi) that are difficult to unequivocally classify even for experts (Polesie et al., 2021; Mevorach et al., 2025). Furthermore, datasets often mix histopathologically confirmed diagnoses (certain) with diagnoses based on visual assessment or follow-up (less certain). Research on the ISIC dataset demonstrates that the quality of Ground Truth (whether derived from a single physician, consensus, or biopsy) has a fundamental impact on algorithm evaluation. An algorithm evaluated against an "imperfect" annotator may be unfairly penalized for a correct diagnosis that conflicts with the provided label (Silva-Clavería et al., 2025; Haggenmüller et al., 2025).

AI models tend to engage in "shortcut learning." If melanoma images in the training set frequently contain rulers (used for measurement prior to excision) while images of benign nevi do not, the network may learn to detect the ruler rather than the tumor. High metrics (*mAP, Accuracy*) in such cases are illusory and will drop drastically on "clean" data. Consequently, reliable evaluation requires the use of external datasets devoid of artifacts, or the application of inpainting techniques to remove these elements prior to analysis (Ünver, Ayan, 2019; Lama et al., 2023).

## Conclusion

The analysis of quality metrics in digital dermatology leads to the conclusion that no single, universal evaluation metric exists. The validation process of AI systems must be multidimensional. In screening tasks (detection), patient safety is the priority, necessitating the maximization of Sensitivity (Recall) to eliminate false negative errors. In surgical planning tasks (segmentation), shape fidelity metrics (Dice, Jaccard, Hausdorff Distance) are crucial. High results reported in the literature (e.g., mAP > 98%) must be interpreted with caution, taking into account the dataset structure and the quality of the Ground Truth.

It is recommended that future research mandatorily report the full confusion matrix and results broken down by individual diagnostic classes to ensure the transparency and safety of deployed solutions.

## References

1. Adebiyi, A., Abdalnabi, N., Smith, E. H., Hirner, J., Simoes, E. J., Becevic, M., Rao, P. (2024). Accurate skin lesion classification using multimodal learning on the HAM10000 and ISIC 2017 datasets. *bioRxiv*. https://doi.org/10.1101/2024.05.30.24308213
2. Alam, T. M., Shaukat, K., Khan, W. A., Hameed, I. A., Almuqren, L. A., Raza, M. A., Aslam, M., Luo, S. (2022). An efficient Deep Learning-Based skin cancer classifier for an imbalanced dataset. *Diagnostics*, *12*(9), 2115. https://doi.org/10.3390/diagnostics12092115
3. Alexandrescu, D. T. (2009). Melanoma costs: A dynamic model comparing estimated overall costs of various clinical stages. *Dermatology Online Journal*, *15*(11), 1. https://doi.org/10.5070/d353f8q915
4. Çetinel, G., Aydın, B. M., Gül, S., Akgün, D., Kara, R. Ö. (2024). An evaluation of skin lesion segmentation using deep learning architectures. *Sakarya University Journal of Computer and Information Sciences*, *7*(3), 449-459. https://doi.org/10.35377/saucis.1543993
5. Crealey, G. E., Hackett, C., Harkin, K., Heckmann, P., Kelleher, F., Lyng, Á., McCarthy, T., McEnery, M., Meaney, C., Roche, D., Tobin, A. (2023). Melanoma-related costs by disease stage and phase of management in Ireland. *Journal of Public Health*, *45*(3), 714-722. https://doi.org/10.1093/pubmed/fdac154

6. Elshahawy, M., Elnemr, A., Oproescu, M., Schiopu, A., Elgarayhi, A., Elmogy, M. M., Sallah, M. (2023). Early melanoma detection based on a hybrid YOLOV5 and ResNet technique. *Diagnostics*, *13*(17), 2804. https://doi.org/10.3390/diagnostics13172804
7. Haggenmüller, S., Wies, C., Abels, J., Winterstein, J. T., Heinlein, L., Nogueira Garcia, C., Utikal, J. S., Wohlfeil, S. A., Meier, F., Hobelsberger, S., Gellrich, F. F., Sergon, M., Hauschild, A., French, L. E., Heinzerling, L., Schlager, J. G., Ghoreschi, K., Schlaak, M., Hilke, F. J., Poch, G. (2025). Discordance, accuracy and reproducibility study of pathologists' diagnosis of melanoma and melanocytic tumors. *Nature Communications, 16*(1). https://doi.org/10.1038/s41467-025-56160-x
8. Khandaker, A. I., Al Shafi, A., Ahmad, M. (2025). Handling Class Imbalance Problem in Skin Lesion Classification: Finding Strengths and Weaknesses of Various Balancing Techniques. *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, 1-6. https://doi.org/10.1109/qpain66474.2025.11171649
9. Lama, N., Hagerty, J., Nambisan, A., Stanley, R. J., Van Stoecker, W. (2023). Skin lesion segmentation in dermoscopic images with noisy data. *Journal of Digital Imaging*, *36*(4), 1712-1722. https://doi.org/10.1007/s10278-023-00819-8
10. Le, P. T., Pham, B., Chang, C., Hsu, Y., Tai, T., Li, Y., Wang, J. (2023). Anti-aliasing attention U-Net model for skin lesion segmentation. *Diagnostics*, *13*(8), 1460. https://doi.org/10.3390/diagnostics13081460
11. Mevorach, L., Farcomeni, A., Pellacani, G., Cantisani, C. (2025). A comparison of skin lesions' diagnoses between AI-Based image classification, an expert dermatologist, and a non-expert. *Diagnostics*, *15*(9), 1115. https://doi.org/10.3390/diagnostics15091115
12. Müller, D., Soto-Rey, I., Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, *15*(1), 210. https://doi.org/10.1186/s13104-022-06096-y
13. Mustafa, W. A., Alquran, H. (2025). Editorial for the special issue "Advances in Medical Image Processing, Segmentation, and Classification". *Diagnostics*, *15*(9), 1114. https://doi.org/10.3390/diagnostics15091114
14. Natha, P., RajaRajeswari, P. (2024). Advancing skin cancer prediction using ensemble models. *Computers*, *13*(7), 157. https://doi.org/10.3390/computers13070157

15. Polesie, S., Sundback, L., Gillstedt, M., Ceder, H., Gyllencreutz, J. D., Fougelberg, J., Backman, E. J., Pakka, J., Zaar, O., Paoli, J. (2021). Interobserver agreement on dermoscopic features and their associations with in situ and invasive cutaneous melanomas. *Acta Dermato-Venereologica*, *101*(10), adv00570. https://doi.org/10.2340/actadv.v101.281
16. Pope, J., Hassanuzzaman, M., Chapman, W., Day, H., Sherpa, M., Emara, O., Adhikari, N., Joshi, A. (2024). *Skin Cancer Machine Learning Model Tone Bias*. ArXiv.org. https://arxiv.org/abs/2410.06385
17. Silva-Clavería, F., Serrano, C., Matas, I., Serrano, A., Toledo-Pastrana, T., Acha, B. (2025). Concordance in Basal Cell Carcinoma Diagnosis. Building a Proper Standard Reference to Train Artificial Intelligence Tools. *Skin Research and Technology, 31*(10). https://doi.org/10.1111/srt.70277
18. Rahi, A. (2025). A hybrid deep learning ensemble for accurate skin cancer classification. *medRxiv*. https://doi.org/10.1101/2025.09.03.25335044
19. Ünver, H. M., Ayan, E. (2019). Skin lesion segmentation in dermoscopic images with combination of YOLO and GrabCut algorithm. *Diagnostics*, *9*(3), 72. https://doi.org/10.3390/diagnostics9030072

## Netography

1. Gevorgyan, A. (2025). *Melanoma cure rate: What patients need to know in 2025*. OncoDaily. Retrieved from: https://oncodaily.com/oncolibrary/melanoma-cure-rate (access date: 10.09.2025).
1. ISIC. (n.d.). *International Skin Imaging Collaboration*. Retrieved from: https://www.isic-archive.com/ (access date: 10.09.2025).
2. Moffat, M. (2024). *Performance metrics in medical image segmentation*. Retrieved from: https://medium.com/@matthewmoffat/different-performance-metrics-in-medical-imaging-segmentation-877f3d20f350 (access date: 10.09.2025).
3. Shah, D. (2023). *Intersection over Union (IoU): Definition, calculation, code*. Retrieved from: https://www.v7labs.com/blog/intersection-over-union-guide (access date: 10.09.2025).
4. Ultralytics. (2025a). *YOLOv8 vs YOLOv5: Evolution of real-time object detection*. Retrieved from: https://docs.ultralytics.com/compare/yolov8-vs-yolov5 (access date: 10.09.2025).

5. Ultralytics. (2025b). *Performance metrics deep dive*. Retrieved from: https://docs.ultralytics.com/guides/yolo-performance-metrics/#how-can-validation-metrics-from-yolo11-help-improve-model-performance (access date: 10.09.2025).
6. Ultralytics. (n.d.). *Mean Average Precision (mAP)*. Retrieved from: https://www.ultralytics.com/glossary/mean-average-precision-map (access date: 10.09.2025).