UDC 004.855.5(045)
DOI:10.18372/1990-5548.82.19373

[1]V. M. Sineglazov,
[2]O. O. Reshetnyk

INTELLIGENT MEDICAL IMAGE PROCESSING SYSTEM USING ZERO-SHOT LEARNING

[1]Aviation Computer-Integrated Complexes Department, Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine
[2]Department of Artificial Intelligence, Educational and Research Institute for Applied System Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
E-mails: [1]svm@nau.edu.ua ORCID 0000-0002-3297-9060,
[2]reshetnyk.oleksii@lll.kpi.ua

***Abstract*—*The work is devoted to the intelligent diagnosis of malignant skin tumors. The classification of malignant skin tumors is presented. The greatest attention was paid to skin melanoma. The modern signs of melanoma were analyzed: Asymmetry, Boundary, Color, and Diameter, and additionally for nodular melanoma: Elevated, Firm, and Growing. A review of works on using artificial intelligence to diagnose malignant skin tumors was performed. A methodology for the intelligent diagnosis of malignant skin tumors was proposed, which is based on the use of preprocessing of dermatoscopic images and solving the segmentation problem based on the use of a hybrid approach, which includes the use of a Segment Anything model based on the combination of the Zero-shot learning model, which consists of an image encoder, prompt encoder, lightweight mask decoder, with YOLOv11. ISIC 2018 was used as the dataset.*

Index Terms—Malignant skin tumors; artificial intelligence; intelligent diagnostics; dermatoscopic images; preprocessing; hybrid approach.

I. INTRODUCTION

The development of artificial intelligence (AI) technologies is changing our lives. Artificial intelligence is gradually being introduced into more and more areas of human activity. Medicine is no exception, where AI helps recognize and analyze various images to assist the doctor in diagnosing.

Oncological diseases have accompanied people throughout history. Cancer is one of the main social, medical, and economic problems of the 21st century. Cancer is the cause of every sixth death of people on earth (16.8%) and the cause of every fourth death (22.8%) that occurs due to non-communicable diseases [1].

Cancer is the cause of death for a large number of people every year, with an estimated 9.7 million deaths in 2022 [1]. The wide variety of malignant tumors, their characteristics, duration of development, and localization add to the difficulties for specialists involved in their definition and characterization. Skin cancer is one of the most common oncological diseases worldwide, and cancer incidence and mortality rates are constantly increasing, mainly in regions with a white population [2]. According to the World Health Organization (WHO), in 2022, about 70,000 people died from non-melanoma skin cancer. The WHO and the ILO (International Labor Organization) have estimated that 1 in 3 deaths from non-melanoma skin cancer per year is caused by work in the open sun [3]. There are 331,647 known cases of melanoma, accounting for 1.7% of all new cancer cases, and 58,645 deaths, accounting for 0.6% of all cancer deaths, in 2022 [1].

Melanoma of the skin is the cause of the majority of deaths from malignant skin neoplasms. Melanoma is a malignant tumor that develops from melanocytes (pigment-forming cells). Melanocytes are cells of non-embryonic origin, located mainly in the basal layer of the epidermis, and produce melanin pigment. Epidermal pigment gives the skin a certain shade and protects it from the effects of ultraviolet radiation.

An ABCD acronym was invented in 1985 [4] to diagnose melanoma. Later, it was expanded to ABCDE. This technique helps to determine with a high probability whether a nevus is dangerous.

An explanation of each word of the acronym is given below.

A – Asymmetry. When one half of the tumor is not the same as the other.

B – Boundary. The border is irregular, jagged, or indistinct.

C – Color. The color varies from one area to another, with a tan or tan-like brown or black hue.

D – Diameter. The diameter of the tumor is greater than 6 mm, which is larger than the size of a pencil eraser.

For nodular melanomas, the following features are defined: ABCDE+EFG [5]:

E – Elevated. New growths that are raised above the skin surface may be suspicious.

F – Firm. The firmness of the growth may be a sign of nodular melanoma.

G – Growing. Nodular melanoma tends to grow. Changes in size may be noticeable over several weeks.

II. Skin Cancer Datasets

Computer analysis of skin lesions typically uses two types of images: dermatoscopic (microscopic) and clinical (macroscopic). Dermatoscopic images allow the examination of features of the lesion that are invisible to the naked eye and are not always available, even to dermatologists. Clinical images are of lower quality but are readily available because they are obtained using conventional cameras.

Dermoscopy is a noninvasive method of obtaining skin images that also allows dermatologists to visualize subcutaneous structures. However, this type of diagnosis has disadvantages because it is highly dependent on the human factor. The accuracy of dermatoscopic diagnosis can vary from 24% to 77% depending on the level of qualification of the dermatologist [6]. Dermoscopy can reduce the level of diagnostic accuracy if used by an inexperienced physician [7].

Therefore, to minimize the probability of errors and avoid false diagnoses, it is extremely necessary to build intelligent systems. Segmentation of skin lesions in images is an important step in achieving this goal. However, the presence of various artifacts (hair or air bubbles), internal factors (variation in the shape and contrast of the lesion), and the variability of image acquisition conditions make segmentation of skin tumors a difficult task.

The lack of images of sufficient quantity and quality is a huge obstacle to the development of segmentation models and effective intelligent systems. Modern machine learning models, including segmentation models, have a huge number of parameters, which allows them to generalize features well when trained on large volumes of labeled data [8]. However, datasets of skin lesions, particularly skin cancer, and all medical image datasets usually have few samples due to the complexity of obtaining and labeling, the right to patient privacy, and the rarity of individual pathologies.

As mentioned earlier, clinical and dermatoscopic images are the most common types of images used to train skin lesion segmentation models. Clinical images help to train the model to segment lesions based on their external features (shape, color, size, edge sharpness). Dermatoscopic images do not capture the surface of the skin but reveal internal skin structures and help identify morphological features (spots, atypical pigment networks, dots/globules, stripes) [9].

The ISIC (International Skin Imaging Collaboration) archive is one of the largest repositories of dermoscopic images. Today, the archive contains 1,156,911 dermoscopic images, 485,127 of which are publicly available. The images are collected from leading clinical centers around the world and obtained using a variety of devices. The involvement of the international community in the input of images is designed to ensure the representativeness of a clinically relevant sample. All images are reviewed to ensure confidentiality and quality. Some of the images were annotated and marked up by skin cancer experts [10].

III. Image Preprocessing

Image preprocessing is an important component of intelligent medical image processing systems because it can improve segmentation results. There are many factors that impair the segmentation of skin lesions, including hair, blood vessels, uneven tumor borders and frames on the image, air bubbles, very small lesions, very large lesions, and low contrast. Preprocessing is designed to reduce the impact of these factors on the performance of the model. The preprocessing operations are listed below.

1. Downsampling. Dermatoscopic images are usually high resolution, i.e., large image size. Most convolutional neural network architectures, such as LeNet, AlexNet, VGG, GoogLeNet, and ResNet, require a fixed input image size (typically 224×224 or 299×299 pixels). Even CNNs that can process images of arbitrary size (e.g., fully convolutional networks) can benefit from downsampling due to reduced computational complexity.

2. Color space transformation. Most models expect images in RGB format, but in some cases, alternative color spaces such as CIELAB, CIELUV, and HSV can be used. Often one or more channels from the transformed space are combined with the RGB channels to improve class resolution, separate luminance and chrominance, ensure invariance to illumination or viewing angle, and remove highlights.

3. Additional input data. In addition to color space transformation, modern work often adds task-specific input data, such as frequency domain representation using discrete Fourier transforms or data based on the physics of illumination and skin imaging.

4. Contrast enhancement. Insufficient contrast is one of the main causes of segmentation errors. If the contrast is insufficient, steps can be taken to pre-enhance the contrast of the images.

5. Color Normalization. Variations in illumination can cause inconsistencies in the segmentation of skin lesions. This problem can be addressed by using color normalization.

6. Artifact Removal. Dermatoscopic images often contain artifacts, the most prominent of which is hair. Hair can be removed before segmentation.

IV. Hybrid Method for Solving the Segmentation Problem

A. Segment Anything Model

Segment Anything (SAM) is a model for image segmentation based on Zero-shot learning, which consists of three components: an image encoder (Fig. 1), a prompt encoder, and a lightweight mask decoder (Fig. 2).

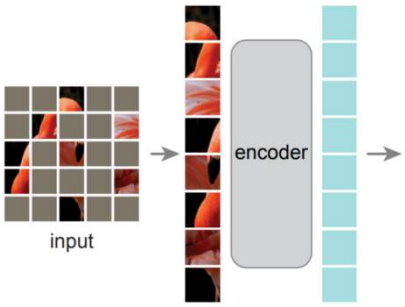

Fig. 1. Architecture of the encoder with Masked autoencoder [11]

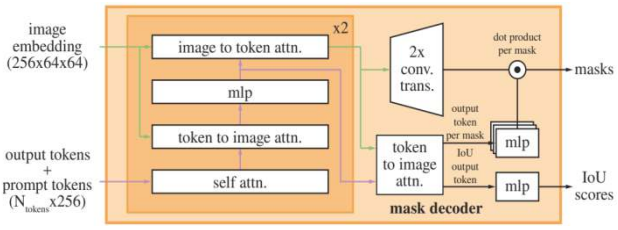

Fig. 2. Architecture of a lightweight mask decoder [12]

The image encoder generates image embeddings of size $C \times H \times W$. Segment Anything employs a Masked Autoencoder (MAE) pre-trained Vision Transformer (ViT-H/16) with windowed attention (14×14) and four evenly spaced global attention blocks for high-resolution images. The input resolution is standardized to 1024×1024 via scaling and padding, resulting in embeddings of size 64×64. These embeddings are downsampled using 1×1 and 3×3 convolutions (256 channels), followed by layer normalization. Computationally expensive operations are minimized by processing each image only once, enabling real-time query handling.

The Prompt Encoder transforms user inputs (prompts) into 256-dimensional embeddings, depending on the type of prompt provided:

- *Points:* Each point is represented by combining its positional encoding (indicating location) with a learned embedding that specifies whether the point belongs to the foreground or background.
- *Boxes:* A rectangular box is represented by two embeddings:

1. The positional encoding of the top-left corner, combined with a learned embedding for the top-left corner.
2. Similarly, the positional encoding of the bottom-right corner combined with a learned embedding for the bottom-right corner.

- *Dense Prompts (e.g., masks):* Dense inputs such as masks are first resized to be one-sixteenth the resolution of the input image. This is done using two 2×2 convolutions with stride 2, which progressively reduce spatial dimensions. The output channels of these convolutions are 4 and 16, respectively. The embeddings are then further processed using a 1×1 convolution to produce a 256-dimensional mask embedding.

If no mask is provided, a learned embedding is added to the image embeddings to indicate the absence of a mask. When text prompts are used, a CLIP-based text encoder is employed, though the approach supports the use of other text encoders as well [13]. This structured process ensures that all prompt types are effectively transformed into a unified embedding space, allowing them to interact seamlessly with the image embeddings in subsequent model components.

The decoder translates images and prompt embeddings into output masks, inspired by Transformer-based segmentation models. It modifies a standard Transformer decoder with learned output tokens.

Each decoder layer performs 4 steps: self-attention for tokens, cross-attention from tokens (as queries) to image embedding, point-wise Multi-Layer Perceptron (MLP) update of each token, and cross-attention from image embedding (as queries) to tokens.

The decoder has two layers, scaling the image embeddings by 4× using two transposed convolutions. Final mask prediction involves an element-wise product of upsampled image embeddings and the MLP output from updated tokens. The Transformer embedding size is 256, with MLPs having an internal dimension of 2048. Cross-attention layers reduce channel dimensions to

128 for efficiency. Attention layers use eight heads. Transposed convolutions for upsampling have 2×2 kernels, strides of 2, and output channels of 64 and 32, with GELU activations and layer normalization.

B. Model YOLOv11

The You Look Only Once (YOLO) framework (Fig. 3) revolutionized the object detection problem by introducing a unified neural network architecture that simultaneously performs bounding box regression and object classification.

The YOLO architecture is based on three fundamental components:

1) Backbone: a core feature extractor that uses convolutional neural networks (CNN) to transform raw data (images) into multi-scale feature maps.

2) Neck: an intermediate processing stage that uses specialized layers to aggregate and improve feature representation at different scales.

3) Head component: a prediction engine that generates final results for object localization and classification based on improved feature maps.

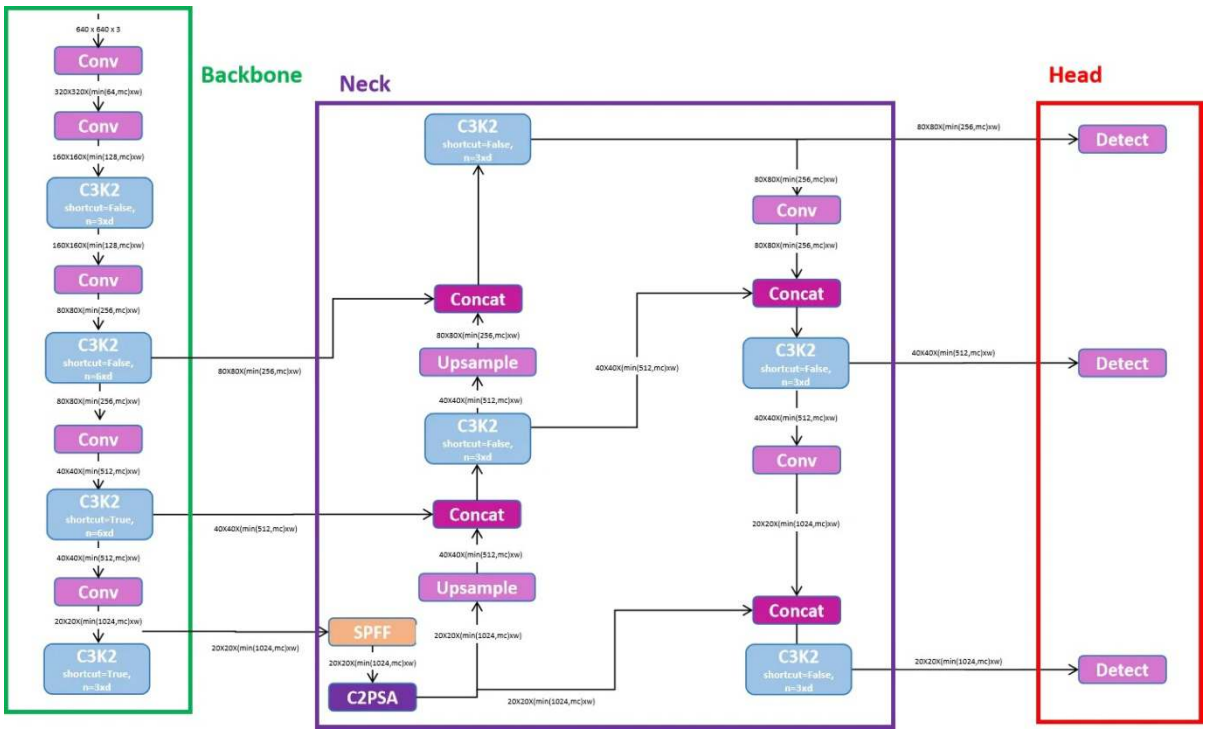

Fig. 3. YOLOv11 Architecture [14]

C. Architecture of the proposed model

The proposed architecture combines YOLOv11 and SAM into a single model (Fig. 4).

The role of YOLOv11 is to identify objects in the photo, i.e. malignant skin tumors. Malignant tumors are regions of interest, which as a result of YOLOv11 work become defined in the photo and are surrounded by a bounding box.

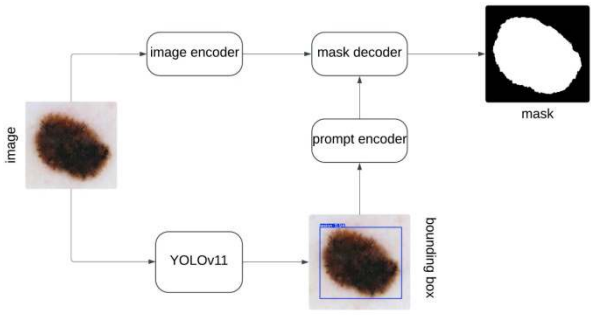

Fig. 4. Architecture of the proposed model

The SAM model in Fig. 1, and Fig. 2, contains an image encoder, a prompt encoder, and a mask decoder. The image and bounding boxes from YOLOv11 are passed to the input of the SAM model, which accepts the frame as a prompt. Thanks to this, SAM "understands" which part of the image should be segmented and performs tumor segmentation. Thus, the model performs the task of medical image segmentation using Zero-shot learning. To improve the results, YOLOv11 was fine-tuned (SAM was frozen), and then SAM (mask decoder) was fine-tuned.

V. Analysis of the Obtained Results

A comparison of the results is given in Table I.

Model_1 is a variant of the proposed model in which YOLOv11 fine-tuning was performed for 100 epochs on 80% of the dataset.

Model_2 is a variant of the proposed model in which YOLOv11 fine-tuning was performed for 100

epochs on 80% of the dataset and SAM fine-tuning was performed for 100 epochs on 80% of the dataset. This model used the same weights for YOLOv11 that were obtained as a result of fine-tuning.

TABLE I. COMPARATIVE TABLE OF MODELS FOR IMAGE SEGMENTATION OF THE ISIC 2018 DATASET

Model	***mIoU***	***Dice***
MedSAM [15]	0.614	0.731
UnSegMedGATc [15]	0.748	0.852
SAM ViT-L BBS5 [16]	–	0.872
SamDSK (HSNet) [17]	–	0.899
Model_1	**0.713**	**0.757**
Model_2	**0.898**	**0.915**

A brief description of the models compared is given below.

1) MedSAM is a SAM-based model trained on a combined medical image dataset containing 1570263 image-mask pairs from 10 modalities, over 30 cancer types, and multiple imaging protocols [18].

2) UnSegMedGATc is an unsupervised model based on pre-trained Dino-ViT [15].

3) SAM ViT-L BBS5 is a SAM model using ViT-L and modifying the bounding box size to 5% of the ground truth bounding box size [16].

4) SamDSK (HSNet) is a model that combines SAM with domain-specific knowledge using an iterative approach that includes training the segmentation model and expanding the annotated dataset.

VI. CONCLUSIONS

An approach to skin tumor segmentation based on the integration of YOLOv11 and SAM models is proposed. The analysis and results demonstrate the effectiveness of this approach, especially after fine-tuning the models on the ISIC 2018 dataset. Two main models were considered:

1) Model_1 – fine-tuning YOLOv11 for 100 epochs on 80% of the dataset: Showed limited segmentation performance, in particular, mIoU = 0.713 and Dice = 0.757, indicating insufficient accuracy in object segmentation due to the lack of SAM adaptation.

2) Model_2 – fine-tuning YOLOv11 and SAM for 100 epochs on the same data: Significantly outperforms Model_1 in all key metrics, achieving mIoU = 0.898 and Dice = 0.915. This highlights the importance of adapting SAM to the specifics of ISIC 2018 data, which significantly improved the quality of segmentation.

The second model demonstrated competitive results compared to other approaches, outperforming them in terms of mIoU and Dice metrics

REFERENCES

[1] F. Bray et al., “Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, Apr. 2024, https://doi.org/10.3322/caac.21834.

[2] Amdad Hossain Roky et al., “Overview of skin cancer types and prevalence rates across continents,” *Cancer Pathogenesis and Therapy*, Aug. 2024, https://doi.org/10.1016/j.cpt.2024.08.002.

[3] “IARC marks Global Non-Melanoma Skin Cancer Awareness Day,” Who.int, 2024. https://www.iarc.who.int/news-events/iarc-marks-global-non-melanoma-skin-cancer-awareness-day (accessed Sep. 22, 2024).

[4] D. S. Rigel, J. Russak, and R. Friedman, “The Evolution of Melanoma Diagnosis: 25 Years Beyond the ABCDs,” *CA: A Cancer Journal for Clinicians*, vol. 60, no. 5, pp. 301–316, Jul. 2010, https://doi.org/10.3322/caac.20074.

[5] “ABCDEFG of melanoma | DermNet NZ,” dermnetnz.org. https://dermnetnz.org/topics/abcdes-of-melanoma (accessed Sep. 22, 2024).

[6] H. Tran, K. Chen, A. C. Lim, J. Jabbour, and S. Shumack, “Assessing diagnostic skill in dermatology: A comparison between general practitioners and dermatologists,” *Australasian journal of dermatology*, vol. 46, no. 4, pp. 230–234, Sep. 2005, https://doi.org/10.1111/j.1440-0960.2005.00189.x.

[7] M. Binder, “Epiluminescence Microscopy,” *Archives of Dermatology*, vol. 131, no. 3, p. 286, Mar. 1995, https://doi.org/10.1001/archderm.1995.01690150050011.

[8] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and Flexible Image Augmentations,” *Information*, vol. 11, no. 2, p. 125, Feb. 2020, https://doi.org/10.3390/info11020125.

[9] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “Diagnostic accuracy of dermoscopy,” *The Lancet. Oncology*, vol. 3, no. 3, pp. 159–65, 2002, https://doi.org/10.1016/s1470-2045(02)00679-4.

[10] “ISIC Challenge,” challenge.isic-archive.com. https://challenge.isic-archive.com/ (accessed Sep. 22, 2024)

[11] K. He, X. Chen, S. Xie, Y. Li, Piotr Dollár, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners,” Nov. 2021, https://doi.org/10.48550/arxiv.2111.06377.

[12] A. Kirillov et al., “Segment Anything,” arXiv (Cornell University), Apr. 2023, doi: https://doi.org/10.48550/arxiv.2304.02643.

[13] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,”

arXiv.org, Feb. 26, 2021. http://arxiv.org/abs/2103.00020

[14] S. N. Rao, “YOLOv11 Architecture Explained: Next-Level Object Detection with Enhanced Speed and Accuracy,” Medium, Oct. 22, 2024. https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speed-and-accuracy-2dbe2d376f71

[15] A. A. Mudit, Shigwan, Saurabh J., and N. Kumar, “UnSegMedGAT: Unsupervised Medical Image Segmentation using Graph Attention Networks Clustering,” arXiv (Cornell University), Nov. 2024, https://doi.org/10.48550/arxiv.2411.01966.

[16] C. Mattjie et al., "Zero-shot Performance of the Segment Anything Model (SAM) in 2D Medical Imaging: A Comprehensive Evaluation and Practical Guidelines," 2023 IEEE *23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, Dayton, OH, USA, 2023, pp. 108–112, https://doi.org/10.1109/BIBE60311.2023.00025.

[17] Y. Zhang, T. Zhou, S. Wang, Y. Wu, P. Gu, and D. Z. Chen, “SamDSK: Combining Segment Anything Model with Domain-Specific Knowledge for Semi-Supervised Learning in Medical Image Segmentation,” arXiv (Cornell University), Aug. 2023, https://doi.org/10.48550/arxiv.2308.13759.

[18] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment Anything in medical images,” *Nature Communications*, vol. 15, no. 1, Jan. 2024, https://doi.org/10.1038/s41467-024-44824-z.

Received October 26, 2024

Sineglazov Victor. ORCID 0000-0002-3297-9060. Doctor of Engineering Science. Professor. Head of the Department of Aviation Computer-Integrated Complexes.
Faculty of Air Navigation Electronics and Telecommunications, National Aviation University, Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine, (1973).
Research area: Air Navigation, Air Traffic Control, Identification of Complex Systems, Wind/Solar power plant, artificial intelligence.
Publications: more than 700 papers.
E-mail: svm@nau.edu.ua

Reshetnyk Oleksii. Master of Computer Science.
Department of Artificial Intelligence, Educational and Research Institute for Applied System Analysis, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine, (2024).
Research Interests: intelligent systems, artificial intelligence, artificial neural networks.
E-mail: reshetnyk.oleksii@lll.kpi.ua

В. М. Синєглазов, О. О. Решетник. Інтелектуальна система обробки медичних зображень з використанням методу нульового навчання

Роботу присвячено інтелектуальній діагностиці злоякісних пухлин шкіри. Представлено класифікацію злоякісних пухлин шкіри. Найбільшу увагу було приділено меланомі шкіри. Проаналізовано сучасні ознаки меланоми: Asymmetry, Boundary, Color, Diameter та додатково для вузлової меланоми: Elevated, Firm, Growing . Виконано огляд робіт з використання штучного інтелекту у діагностиці злоякісних пухлин шкіри. Запропоновано методологію інтелектуальної діагностики злоякісних пухлин шкіри, яка базується на використанні попередньої обробки дерматоскопічних зображень та розв’язанні задачі сегментації на основі використання гібридного підходу, який включає застосування Segment Anything model на основі об’єднання моделі Zero-shot learning, яка складається з image encoder, prompt encoder, lightweight mask decoder з YOLOv11.В якості датасету було використано ISIC 2018.

Ключові слова: злоякісні пухлини шкіри; штучний інтелект; інтелектуальна діагностика; дерматоскопічні зображення; попередня обробка; гібридний підхід.

Синєглазов Віктор Михайлович. ORCID 0000-0002-3297-9060. Доктор технічних наук. Професор. Завідувач кафедри авіаційних комп’ютерно-інтегрованих комплексів.
Факультет аеронавігації, електроніки і телекомунікацій, Національний авіаційний університет, Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна, (1973).
Напрям наукової діяльності: аеронавігація, управління повітряним рухом, ідентифікація складних систем, вітроенергетичні установки, штучний інтелект.
Кількість публікацій: більше 700 наукових робіт.
E-mail: svm@nau.edu.ua

Решетник Олексій Олександрович. Магістр комп’ютерних наук.
Кафедра штучного інтелекту, Навчально-науковий інститут прикладного системного аналізу, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.
Освіта: Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, (2024).
Напрям наукової діяльності: інтелектуальні системи, штучний інтелект, штучні нейронні мережі.
E-mail: reshetnyk.oleksii@lll.kpi.ua