INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

IJIE

Vol. 16 No. 1 (2024) 293-300
https://publisher.uthm.edu.my/ojs/index.php/ijie

An Optimized Semantic Segmentation Framework for Human Skin Detection

Audrey Huong[1]*, Xavier Ngu[1]

[1] *Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, MALAYSIA*

*Corresponding Author: audrey@uthm.edu.my
DOI: https://doi.org/10.30880/ijie.2024.16.01.024

Article Info

Received: 27 October 2023
Accepted: 14 February 2024
Available online: 22 May 2024

Keywords

Skin detection, segmentation, PSO, AlexNet, Jaccard

Abstract

The study incorporating optimization strategy in semantic segmentation is underexplored in dermatology. Existing approaches used complex and various heuristic designs of image processing algorithms and deep models customized for skin detection problems. This paper demonstrates Particle Swarm Optimization (PSO)-incorporated AlexNet framework for the skin segmentation task. The results from testing the trained model are promising. The model produced satisfactory performances even with a strict split of 50 %, confirming the high efficiency of the proposed framework. The mean Jaccard index and Dice similarity measures evaluated between the annotated and predicted mask ranged from 0.80 to 0.93 in the binary classification of pixels as "skin" versus "background". This work identified that the location and color variability of skin pixels in the training data are crucial to obtaining a good skin segmentation performance. Further works that can be explored in this area include adopting a robust preprocessing strategy to increase data variability and improve model generalization or implementing an optimization-enhanced strategy on the existing segmentation models for comparison.

1. Introduction

Skin is a multi-layered tissue consisting of stratum corneum, epidermis, deeper connective dermis tissue layer, and subcutaneous adipose. It is the largest organ in the human body and serves as a strong barrier to protect the internal organs from external pathogens, harmful rays, and harsh environments, making it an immune-competent tissue. A disrupted skin barrier integrity is often associated with inflammatory and many immune-mediated diseases, so skin examination is the first step in the clinical examination of a patient. It is also a key characteristic to locate and track humans. Accurate and efficient human skin detection and segmentation is a challenging yet important task in many applications, such as healthcare support, security and surveillance, automation and interactive communications, and nutritional industries. Some of the important works in the past include the detection of facial micro-expressions in coma patients [1] and neurological impaired and dementia patients [2]. There is also a growing interest in this research area and telemedicine for automatic diagnosis of different skin conditions [3-4] and remote monitoring for elder care. Researchers in [5] and [6] used skin detection to detect explicit images and pornography content. Skin is also a powerful feature for pedestrian detection in autonomous driving systems, multi-person tracking, crowd density estimation [7], and privacy protection through real-time surveillance by differentiating humans and machines. Xie et al. [8] investigated the relationship between the diversity of the detected skin tone and marketing communication in the fashion industry. Other works adopted skin detection and segmentation for noncontact continuous monitoring of vital signs [9], sign language

recognition, and skin care product recommendations. Most conventional systems use infrared and thermal imaging technology to detect skin. Meanwhile, advancements in the technologies using multispectral imaging and Terahertz attenuation reflectance technologies offer the potential for further improvement in imaging sensitivity.

Recent years have seen a remarkable rise in the number of innovations invented and efforts made involving artificial intelligence (AI) in these industries. The emergence of this technology brings a new air to society by promoting a better quality of life, offering better healthcare and opportunity, simplifying a used-to-be laborious task, and providing an efficient and effective system for equal-balanced social life. Even though the trade-offs of introducing AI in most everyday tasks have been a hot debate, the advantages of this technology in terms of low effort and high efficiency cannot be ignored. They are transforming how humans live, work, and interact with others. AI is particularly pronounced in medical image analysis to assist clinical decision-making and speed up the diagnosis process where early treatment can be initiated. AI-enabled skin detection and segmentation technologies are currently used in medical and dermatological research for remote monitoring and classification of skin lesions and diseases [10], jaundice detection, rehabilitation assessment, early detection of diabetic foot ulcers, and aesthetics class prediction [11]. Among these applications, skin disease classification is the major medical complication in dermatology, affecting one-third of the world population [12]. Misdiagnosis of disease, especially malignant cancers, may lead to late treatment and increased morbidity.

Semantic segmentation is a critical step for the skin disease diagnosis task. Recent studies in the field include comparing various feature extractor techniques for skin detection and strategies to improve the designed system's detection accuracy and segmentation performance. Since human skin has distinct physical and color properties that allow for its differentiation from nonskin objects, texture-based techniques and color models, i.e., RGB (red, green, blue), HSV (hue, saturation, value), and YCbCr (Luma and chroma components of an image), are used as a feature extractor for skin detection and localization using machine learning methods [13], convolutional neural networks (CNN), statistical methods (e.g., histogram and distribution analysis), and traditional methods, such as wavelet transform [14].

Deep learning methods using a pretrained CNN, for example, AlexNet, VGGNet, and ResNet, or their variants, are an increasingly popular approach. However, these methods have not been rigorously studied and tested in dermatology. Existing Fully-CNNs, such as U-Net and SegNet, are available for semantic segmentation, but failed to process complex feature information [15]. These models transferred learning for the problem using large amounts of labeled data to learn and extract important features in the input images automatically. The ensemble system that combined CNN with map expansion technique and outer residual skip connection-based deep CNN are alternatives that have been shown to work considerably well for skin segmentation [16-17]. However, these methods exhibit heavy computational burdens and tedious design processes. Besides, adapting these models to the target dataset depends heavily on the hyperparameters used in the training. Thus, careful tuning of these parameters is required to infer the most relevant features and enrich the learning of quality representations. Even though optimization techniques, such as genetic algorithm (GA), PSO, and Bayesian algorithm (BA), are among the available approaches, grid search that requires minimal programming effort is most often used for the problem. The process can be laborious with a high probability of converging to local minima. Instead of using a complex feature engineering method, this paper demonstrates the PSO auto-search method to address the optimization problem and improve skin segmentation efficiency using a simple network modified from AlexNet. Section 2 presents the data used and introduces the proposed framework, followed by the presentation of the results in Section 3. Section 4 discusses results from the model trained with different experimental settings and outlines future directions before concluding in section 5.

2. Material and Methods

2.1 Human Skin Dataset and Data Handling

There are many human skin datasets publicly available for use; for demonstration of the proposed framework, this study used a private face and skin dataset obtained from the Visual and Audio Signal Processing Lab (VASP) University of Wollongong (https://documents.uow.edu.au/~phung/download.html) due to the availability of its ground-truth. This dataset contains 4,000 RGB images with diverse background scenes, lighting conditions, and skin types. The annotated skin-segmented images contain exposed skin regions, e.g., facial skin, arms, hands, neck, and background (i.e., nonskin) pixels. The size of these images is 352 × 288 pixels; an image resizing process was performed to change the size to 300 × 400 pixels to match the input of the proposed model. Shown in Fig. 1 are examples of the resized images and the binary ground-truth images (pixel value "0": skin and "1": nonskin). These images were divided into training, validation, and testing sets using two split ratio settings of 0.70/0.10/0.20 (i.e., 70 % split) and 0.50/0.05/0.45 (i.e., 50 % split). A constant random seed number was used in the splitting process for the reproducibility of the results. For a more objective evaluation of the efficiency and robustness of the proposed system, this study has not considered other performance improvement strategies, such as dataset enrichment through image augmentation or synthetic data generation and regularization methods.

Int. Journal of Integrated Engineering Vol. 16 No. 1 (2024) p. 293-300

Fig. 1 *(Left) Examples of human skin images and (Right) the annotated images*

2.2 Skin Semantic Segmentation Network

AlexNet is an efficient CNN network for image recognition problems. It contains series and continuous modules comprised of convolutional layers; thus, it can easily be modified for different operations compared to the GoogleNet variants that contain inception modules and ResNet with residual connections. Meanwhile, VGG variants are heavy and computationally exhaustive. In the pre-experiment simulations, the initial performance of the VGGNet and GoogleNet have been explored, but both networks failed to converge satisfactorily using the considered dataset. The VGG models, even with the shallow VGG-16, took about ten times longer than AlexNet in its training and optimization process. So, only the results from AlexNet are reported in this paper.

Fig. 2 shows the AlexNet architecture modified for the segmentation task. A large input size of 300 × 400 pixels is used in this study to refine the segmented results, and a padding of 100 was applied on the first convolutional layer of the model (Conv1) to prevent mismatch problems. Two-dimensional convolutional layers (Conv6 and Conv7) with 4,096 neurons have been added after the POOL5 max-pooling layer for efficient inference and learning. The output feature from Conv7 is fed into the score layer to obtain a feature map of the semantic labels. These feature maps are upsampled using a transposed convolutional layer (deConv) before the output is cropped to align with the input size. This is followed by pixels' label prediction in the Softmax layer.

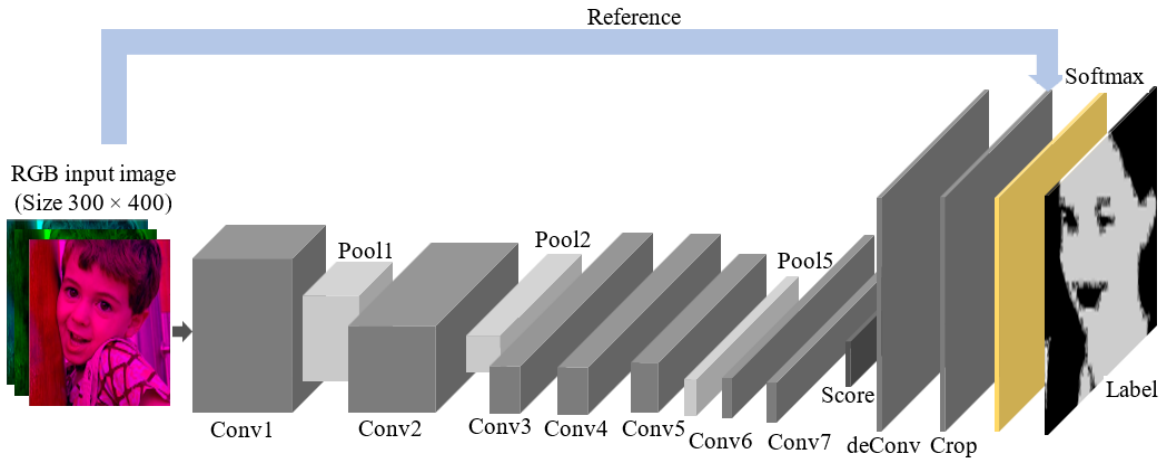


Fig. 2 *Skin segmentation AlexNet architecture*

2.3 PSO Optimization Framework

This paper adopts PSO to find the optimal solution for the training hyperparameters problem due to its fast convergence rate compared to other techniques [18]. Training solver type, epoch number, mini-batch size, and initial learning rate were chosen as the hyperparameters to be optimized due to their significant effect on the model learning efficiency. The objective function to be minimized is shown in Eq. (1).

$$f(T_{acc}, V_{acc}, t) = (100 - T_{acc})^2 + (100 - V_{acc})^3 + {}^{t}/_{1000} \quad (1)$$

T_{acc}, V_{acc}, and t represent training and validation accuracies and training time, respectively. This process begins with randomly initializing 20 particles in the search space defined in Table 1. While these ranges were mainly chosen to meet the limitation of the GPU memory, other factors that help in the decision are (1) based on the

recommendation of the relevant study [18] for mini-batch size and initial learning rate, while (2) maximum epoch number is empirically decided during the pre-experiment runs, by weighing the computational efficiency and classification performance. The performance of each particle (i.e., candidate solution) is evaluated in Eq. (1). The current best solution is identified before each particle changes its position and velocity in the iteration that follows. After each iteration, the current best solution is updated to the new one. The process is repeated five times before determining the global optimal hyperparameter solution. The termination criteria include when the maximum epoch number is reached or if the V_{acc} has not increased in the last ten consecutive evaluations. The chosen hyperparameter set is used in training the model for segmenting the unseen testing images. This process has been implemented on the dataset of different splits.

Table 1 *The boundary limit of the considered hyperparameters in the search space*

Hyperparameter	Lower limit	Upper limit
Solver, L	1 →3: {*Adam*, *Sgdm*, *RMSProp*}	
Epoch number, α	50	100
Mini-batch size, β	8	64
Init. learning rate, γ	$1e^{-5}$	$1e^{-1}$

2.4 Image Segmentation Performance

The predicted two-class segmented mask consists of an image containing pixels with values "0" and "1", representing that of the skin and non-skin region, respectively. Thus, similarity and overlap percentages between the predicted (*PRED*) and ground-truth binary map (*GT*) are used to evaluate the performance of the system training for the skin segmentation problem. The Intersection Over Union (IU or Jaccard index) and Dice similarity (DS) in Eqs. (2) and (3) are used as performance measures. IU is the probability of correctly classifying pixels of a class, while DS determines the area of overlap divided by the total pixel number in both *PRED* and *GT* masks. Their value can range from 0 to 1, with higher scores indicating better segmentation quality. This study also considered pixel accuracy (ACC) as a prediction correctness metric in Eq. (4). It is the ratio of correctly classified pixels to the total number of pixels.

$$\text{IU} = \frac{Area\ of\ Overlap\ (I)}{Area\ of\ Union\ (U)} \tag{2}$$

$$\text{DS} = 2\frac{Area\ of\ Overlap\ (I)}{GT+PRED} \tag{3}$$

$$\text{ACC} = \frac{TP+TN}{TP+TN+FN+FP} \tag{4}$$

TP and *TN* are the number of pixels correctly identified as skin and nonskin class, respectively. *FP* represents the incorrect prediction of nonskin pixel number as the skin class, while *FN* is the opposite.

3. Results and Analysis

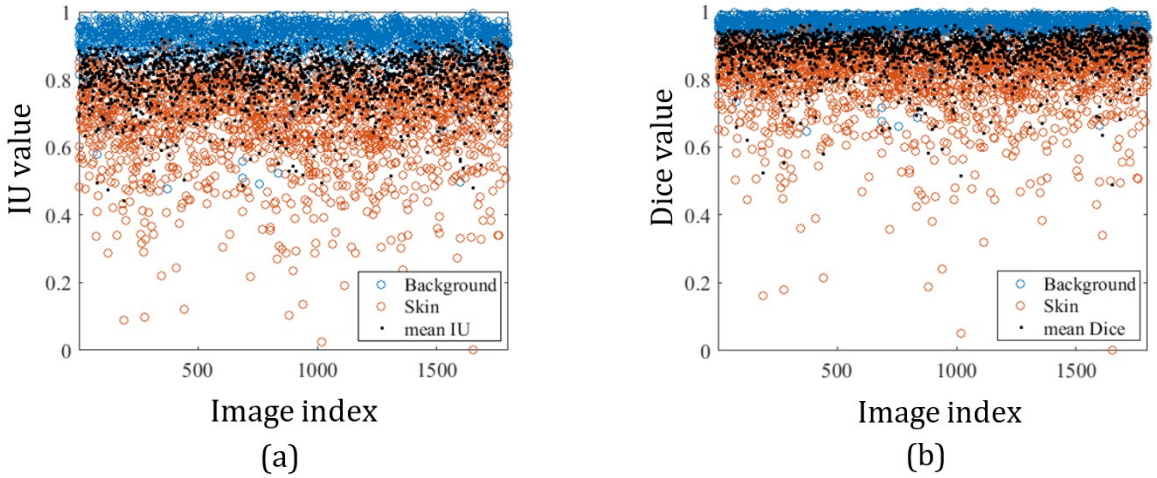
The optimal hyperparameter setting for training the AlexNet model in Fig. 2 for the skin segmentation problem is identified as {L = '*Sgdm*', α = 69, β = 62, and γ = 0.054} and {L = '*Sgdm*', α = 100, β = 57, and γ = 0.0809} using the dataset of 70 % and 50 % split, respectively. The mean search time is 1,263 s and 1,380 s, respectively, on an NVIDIA Tesla K80 GPU. The final network is 27 layers deep with 56.8 million parameters for training. The model trained with 70 % of data samples performed slightly better than the 50 % split. So, results from this rigid split are presented in more detail in the following. Fig. 3 shows the IU and DS values calculated for skin and nonskin (background) regions. Table 2 shows their mean value averaged from the results for both classes. Also shown in this table is the average ACC. There is a noticeable issue with the class imbalance in most testing images, wherein the number of nonskin pixels (i.e., background) in an image is considerably larger than the target pixels (skin region). Thus, in Table 2, this work addressed the problem by including mean weighted IU as an important performance metric. The best and the worst-performing test images identified based on the mean IU and DS results in Fig. 3 were chosen for further investigation in Figs. 4 and 5. Also shown in the figures are the original image and the overlaid mask of the predicted classes. The right of these figures shows the gradient-weighted class activation mapping (grad-CAM) of the deConv output features.

Int. Journal of Integrated Engineering Vol. 16 No. 1 (2024) p. 293-300

Table 2 *Evaluated performance metrics of model trained using different data splits*

Split setting (Train/Val./Test)	Evaluated metrics*			
	Mean IU	Mean DS	Mean ACC	Weighted IU
70 % split (0.7/0.1/0.2)	0.82	0.89	0.94	0.93
50% split (0.5/0.05/0.45)	0.80	0.88	0.93	0.92

*IU: Intersection of Union, DS: Dice similarity, ACC: Pixel accuracy



(a) (b)

Fig. 3 *The calculated (a) Intersection of union (IU) and (b) Dice similarity (Dice) evaluated for each testing image. These results are obtained from the experiment using the rigid 50 % split*

Fig. 4 *(Left) The best-performing test image and the overlaid predicted map (yellow hue) with mean IU and DS of 0.93 and 0.88, respectively. (Right) The deConv layer feature activation map*

Fig. 5 *(Left) The worst-performing test image and the overlaid predicted map (yellow hue) with mean IU and DS of 0.48 and 0.49, respectively. (Right) The deConv layer features activation*

4. Discussions

The most conventional method for skin detection requires tedious manual labor annotation, while the automatic approach using the machine learning method involves the time-consuming and meticulous design of a model for the problem. This study overcomes this challenge by using an automatic search algorithm in the training process to optimize the model learning efficiency and improve its generalization ability using the relatively simple and shallow network in Fig. 2.

The good performance of this model trained with a 70 % split in Table 2 is due to the rich features representation of input images learned by the network. On the contrary, the experiment using a strict 50 % split deprived the model of a significant proportion of the dataset, setting aside large and different testing data for evaluating the model. This experiment design is closer to real-world scenarios when a model trained with limited training data is used to predict extensive unseen data. The comparable performance between the different split ratios in the evaluated metrics in the table confirmed the feasibility of the proposed strategy.

The IOU and DS metrics evaluated for each pixel in the testing set from the 50 % split experiment in Figs. 4 and 5 showed a comparative inferiority in the model's recognition of skin pixels. The mean IU and DS evaluated for the skin pixels are given by 0.7 and 0.8, respectively, implying that, on average, 30 % of the skin region has been misclassified as the background. Conversely, the background pixels have near-perfect prediction accuracy, with IU and DS ranging between 0.9 and 1.0. The areas covered by skin in most testing images are noticeably smaller than the background. The weighted IU metrics considering the number of pixels for each class showed a good performance of 0.92 in Table 2 due to the accurate prediction of the extensive background pixels. The same reason applies to the high mean pixels' accuracy of 0.93 in Table 2. Two main reasons have been identified for the misclassification of the skin pixels. The first is the limited variability in the location of the target pixels in training images. In nearly all the images, the object (human) is located near the photograph's center, resulting in the hot regions (red color) crowded around the center of the activation maps, e.g., shown on the right of Figs. 4 and 5. These high-heat regions indicate the network's focus on skin prediction; thus, the model is not expected to work equally well when the object is located at the edge of the image. The second is the skin color factor. The model has been trained to recognize mostly the skin of Caucasians, Mediterraneans, and Asians. These are the populations with lighter skin colors with Fitzpatrick skin phototype I-IV. Therefore, darker skin tones (type V-VI), such as those of African or Caribbean origin in Fig. 5, may be harder to detect. The image shows a small number of sling bag pixels whose color is similar to the skin and has been misclassified as "skin" (circled in red). This produces a very high IU and DS score (about 0.98) for background pixels but zero overlapping between the segmented and annotated skin regions (IU and DS = 0) in Fig. 3. Other factors that influence skin detection accuracy include the visual (size, shape, and connectivity) components in the image. The model can work well in most images with high textural details and good contrast between skin regions and the background, such as the high-contrast image in Fig. 4.

Although a larger training set has been shown in Table 2 to improve the detection accuracy of skin pixels, the inclusion of data enlargement strategies using different image preprocessing techniques, such as data augmentation and generation, is expected to enhance the model performance further. The augmentation that involves affine transformations would also modify the deConv gradient-weighted activation maps in Figs. 4 and 5, allowing better skin recognition at all points in an image. The deeper models, namely GoogleNet and VGG-16, have been investigated in the pre-experimental stage. However, the initial results showed overfitting problems. Therefore, other directions worth exploring include truncating these models to reduce the architecture complexity using the proposed system as the benchmark. The original AlexNet is not designed for the problem explored in this study, causing insufficiency of the model even with the larger dataset in Table 2. Hence, future comparisons to U-Net and SegNet models that are popular for various segmentation tasks would also be valuable.

5. Conclusion

This research demonstrates the use of the PSO method in training the AlexNet modified for the skin semantic segmentation task. The results comparing the difference in the segmentation performances of the model trained using 70 % and 50 % splits showed the potential of the proposed framework for automatic fine-tuning of important hyperparameters for fast and optimal convergence in the model training. This approach is notably more practical and efficient than the prior works that required the sophisticated design of the model structure. The results showed an overall satisfactory matching between the predicted and annotated masks, with the evaluated mean IU and DS ranging from 0.80 to 0.93. In the future, data enhancement and image preprocessing algorithms can be implemented to improve the system's performance further. The new and improved system can facilitate future comparison with other popular segmentation networks.

Acknowledgement

This work was partially supported by Universiti Tun Hussein Onn Malaysia (TIER1 Q381) and the Ministry of Higher Education Malaysia through Fundamental Research Grant Scheme (FRGS/1/2020/TK0/UTHM/02/28).

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design, original draft preparation, methodology, software, investigation:*** *AH;* ***validation, writing reviewing and editing, funding acquisition:*** *XN. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Wang, H., Huang, J., Wang, G., Lu, H., & Wang, W. (2022). Surveillance Camera-based Cardio-respiratory Monitoring for Critical Patients in ICU. *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. doi: 10.1109/BHI56158.2022.9926954.

[2] Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020). The modeling of human facial pain intensity based on Temporal Convolutional Networks trained with video frames in HSV color space. *Applied Soft Computing*, 97, pp. 1-14. https://doi.org/10.1016/j.asoc.2020.106805

[3] Nyemeesha, V., & Ismail, B. M. (2021). A Systematic Study and Approach on Detection of Classification of Skin Cancer Using Back Propagated Artificial Neural Networks. *Turkish Journal of Computer and Mathematics Education*, 12, (11), pp. 1737-48.

[4] Alwakid, G., Gouda, W., Humayun, M., & Sama, N. U. (2022). Melanoma Detection Using Deep Learning-Based Classifications. *Healthcare (Basel)*, 10 (2481), pp. 1-18. DOI: 10.3390/healthcare10122481

[5] Subaeki, B., Gerhana, Y. A., Rusyana, M. B. K., & Manaf, K. (2023*).* Digital Image Processing Using YCbCr Colour Space and Neuro Fuzzy to Identify Pornography. *Jurnal Online Informatika*, 8 (1), pp. 122-130. https://doi.org/10.15575/join.v8i1.1070

[6] Albahli, S. (2022). Transfer Learning on Deep Neural Networks to Detect Pornography. *Computer Systems Science & Engineering*, 43, pp. 1-17. https://doi.org/10.32604/csse.2022.022723

[7] Kim, K.R., Koh, Y.J., & Kim, C. S. (2020). Instance-Level Future Motion Estimation in a Single Image Based on Ordinal Regression and Semi-Supervised Domain Adaptation. *IEEE Access*, 8, pp. 115089-115108. DOI: 10.1109/ACCESS.2020.3003751

[8] Xie, W., Overgoor, G., Lee, H. H., & Han, Z. (2023). Automated Detection of Skin Tone Diversity in Visual Marketing Communication. *Proceedings of the 56th Hawaii International Conference on System Sciences*. pp. 3817-3827.

[9] Khanam, F. T. Z, Perera, A. G., Al-Naji, A., Gibson, K. & Chahl, J. (2021). Non-Contact Automatic Vital Signs Monitoring of Infants in a Neonatal Intensive Care Unit Based on Neural Networks. *Journal of Imaging*, 7, pp. 1-19. https://doi.org/10.3390/jimaging7080122

[10] Nisar, H., Ch'ng, Y. K., & Ho, Y. K. (2020). Automatic Segmentation and Classification of Eczema Skin Lesions Using Supervised Learning. *2020 IEEE Conference on Open Systems (ICOS)*. doi: 10.1109/ICOS50156.2020.9293657

[11] Chen, M., Zhou, P., Wu, D., Hu, L., Hassan, M.M., & Alamri, A. (2020). AI-Skin: Skin disease recognition based on self-learning and wide data collection through a closed-loop framework. *Information Fusion*, 54, pp. 1-9. https://doi.org/10.1016/j.inffus.2019.06.005

[12] Karimkhani, C., Dellavalle, R.P., Coffeng, L.E., et al. (2017). Global skin disease morbidity and mortality: an update from the Global Burden of Disease Study 2013. *JAMA Dermatol.*, 153(5), pp. 406– 12. doi:10.1001/jamadermatol.2016.5538

[13] Moumene, M.E., Benkedadra, K., & Berras, F. Z. (2022). Real Time Skin Color Detection Based on Adaptive HSV Thresholding. *Journal of Mobile Multimedia*, 18, pp. 1617–1632. https://doi.org/10.13052/jmm1550-4646.1867

[14] Pilania, U., Tanwar, R., & Gupta, P. (2022) An ROI-based robust video steganography technique using SVD in wavelet domain. *Open Computer Science*, 12, pp. 1-16.

[15] Zhao, Y., Feng, G., Wang, Y., Wang, X., Wang, Y., Lu, H., Xu, W., & Wang, H. (2022). A new algorithm for intelligent detection of geohazards incorporating attention mechanism. *International Journal of Applied Earth Observation and Geoinformation*, 113, pp. 1-13. https://doi.org/10.1016/j.jag.2022.102988

[16] Lee, J., Kim, E., Lee, S., Lee, J., & Yoon, S. (2019). FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1- 12. https://doi.org/10.48550/arXiv.1902.10421

[17] Arsalan, M., Kim, D.S., Owais, M., & Park, K.R. (2020). OR-Skip-Net: Outer residual skip network for skin segmentation in non-ideal situations. *Expert Systems with Applications*, 141. https://doi.org/10.1016/j.eswa.2019.112922
[18] Huong, A., Tay, K.G., Gan, K.B., & Ngu, X. (2022). Hierarchical Optimization Framework for Pigmented Lesion Diagnosis. *CAAI Transactions on Intelligence Technology*, 7 (1), pp. 34-45. https://doi.org/10.1049/cit2.12073.