

Overcoming Challenges in Detecting Multi-class Skin Cancer from Small Dataset Using Transfer Learning Approach

Abstract—Automatically detecting different skin cancers from an image of a skin lesion can greatly help medical professionals in early diagnosis. It can also aid in non-invasive skin cancer identification. However, the lack of dataset availability, bias, class imbalance and suitable ways to work with a small sample of data are poorly defined areas for skin cancer identification. This research addresses these issues with a novel algorithm incorporating tuning image augmentation and training YOLOv8 in a transfer learning manner. A dataset with 1000 images of five different skin cancers was collected and annotated for cancer detection with the YOLOv8 algorithm. The base YOLOv8 was first trained for a larger epoch to create a baseline model, and the trained model was stripped to retrain with optimized hyperparameters. The final model performed closely to other models trained on large datasets. Evaluating unseen images with our trained model confirmed its applicability to real-life scenarios. This study demonstrates the effectiveness of using this method in improving the YOLOv8’s detection performance and providing a more effective solution to multi-class skin cancer detection from small sample sizes.

Index Terms—Small Dataset, YOLOv8, Transfer Learning, Multi-class Skin Cancer

I. INTRODUCTION

One of the most pervasive types of cancer worldwide is skin cancer. Skin cancer arises in various kinds including squamous cell carcinoma, intraepithelial carcinoma, melanoma and basal cell carcinoma [1]. The three tissues composing human skin are the dermis, epidermis and hypodermis [2]. Melanocytes found in the epidermis portion can produce melanin under certain conditions at a highly exceptional place. Melanin, for example, is produced after consistent exposure to intense UV Light from sunshine [3]. Melanoma is a malignant type of skin cancer that is caused by the unusual growth of melanocytes [4]. A key factor in improving patient outcomes and lowering death rates in skin cancer is early detection and accurate categorization. Deep learning and computer vision algorithms have demonstrated favourable results in the classification and localization of affected skin regions in recent years, providing automated and effective analysis of dermatological pictures [5]. Researchers want to improve the accuracy and speed of skin cancer treatment by utilizing computer vision algorithms, which will allow for prompt interventions and individualized treatment plans.

Nonetheless, there are still several difficulties in using computer vision to classify skin cancer. The challenge for research is to create a reliable algorithmic pattern that can distinguish between malignant and benign skin regions with

high accuracy while considering a variety of characteristics, including handling complicated skin lesion patterns and lesion size, shape, texture, and colour [6]. The wide range of skin tones, ethnicities, and imaging conditions also makes classification extremely difficult. Image analysis and machine learning techniques are frequently employed in computer vision-based skin cancer detection research for bettering the accuracy, efficacy and accessibility of skin cancer therapy and diagnosis [7]. This field contains several significant study domains.

The first area of emphasis is affected skin lesion identification and segmentation, which precisely locates and defines skin lesions in pictures to facilitate further examination and research. Researchers also strive to create effective techniques for feature extraction and representation [8]. Moreover, very accurate risk assessment and categorization systems are produced by combining machine learning and deep learning approaches. Using visual analysis to combine clinical data—such as patient history, demographics, and genetic information—with visual analysis is another area of research that aims to increase diagnostic precision and provide a thorough knowledge of the features of skin lesions [9].

Object Detection tasks evolved with the introduction of deep learning. In the field of computer vision, according to each requirement, YOLO has undergone multiple modifications since its launch, with each iteration bringing enhanced accuracy and speed. YOLO’s end-to-end approach estimates bounding boxes and class probabilities directly, streamlining the object detection workflow [10]. It is simpler to implement and deploy because of its simplicity, which also contributes to speeding up training and inference times. YOLO effectively strikes a balance of computing efficiency and precision. YOLO simplifies region proposal networks by analyzing the entire image in a single pass, thus minimizing the amount of memory and computation needed [11]. Multiclass skin cancer detection tasks can be object detection based but the availability of an annotated dataset and its size are two of the prime obstacles to getting the desired outcome. These issues are commonly visible in the domain of medical imaging.

In this work, we attempted to mitigate these issues by using the state-of-the-art YOLOv8 algorithm through fine-tuning and transfer learning approach to achieve multiclass malignant skin lesion detection results. With a satisfactory accuracy rate on a test set, our ensembled algorithms identified all five types of cancer. The data augmentation techniques implemented assisted YOLO in achieving this success on unseen images.

II. LITERATURE REVIEW

At the early time of skin cancer detection through machine learning, classification tasks primarily relied on optical coherence tomography (OCT). Following OCT imaging, preprocessing steps were applied, focusing primarily on binary classification through accuracy and precision assessments. Before 2016, machine learning was primarily used for classification tasks, e.g.; binary classification [12]. Then, the International Skin Imaging Collaboration (ISIC) began a challenging competition on skin cancer detection in 2016, becoming a pivotal repository for researchers. They started to use convolutional neural networks (CNNs) to classify skin cancer and malignancy detection by medical image analysis. Subsequently, with CNNs, researchers also utilized various models like Deep CNNs, SVMs, AlexNet, ResNet-18, and VGG16, etc on the ISIC dataset. Tanna R et al. proposed binary classification methods for melanoma skin cancer detection, employing SVM and CNN techniques with feature extraction through image processing [13]. Hekler et al. studied 112 dermatologists who independently assessed the same images to classify multiclass classification of skin lesions. Using CNN's with gradient boosting method, they achieved higher accuracy (82.95%) and sensitivity (89%) [14]. Later, Ali K et al. focused on multiclass skin cancer classification using convolutional neural networks (CNNs) with transfer learning on Efficient-Nets B0-B7 [15]. Mahbod A et al. investigated how image size affects skin lesion classification using pre-trained CNNs. They proposed an MSM-CNN approach that achieves a balanced multi-class accuracy of 86.2% [16]. Kassem MA et al. proposed transfer learning and GoogleNet, achieving high classification accuracy, sensitivity, specificity, and precision [17]. In recent times, yolo emerged as a very popular model for skin cancer multiclass classification. It has some advantages over CNN like less number of background errors, real-time speed while maintaining high average precision etc. Singh SK et al. made a model that used YOLO V3 architecture based on ISIC 2017 and ISIC 2018 datasets [18]. Later, Aishwarya N et al. proposed a deep neural network-based approach using YOLO V3 and V4, with mean average precision scores of 88.03% and 86.52% for YOLO V3 and V4, respectively, across 4389 images [12].

But YOLO older versions (V1, V2) have different limitations like difficulty detecting small objects, high localization error, fixed number of predictions, low speed etc. The latest version of YOLO, for example, yolo V8 addresses these limitations by boasting high accuracy and faster speed. Based on these previous studies it is seen that multiclass classification is not thoroughly researched using YOLO. By inspiring from those classifications and works on multi-class classification we use the latest version of Yolo v8 which is more improved from the previous version of YOLO for the detection of multi-class classification.

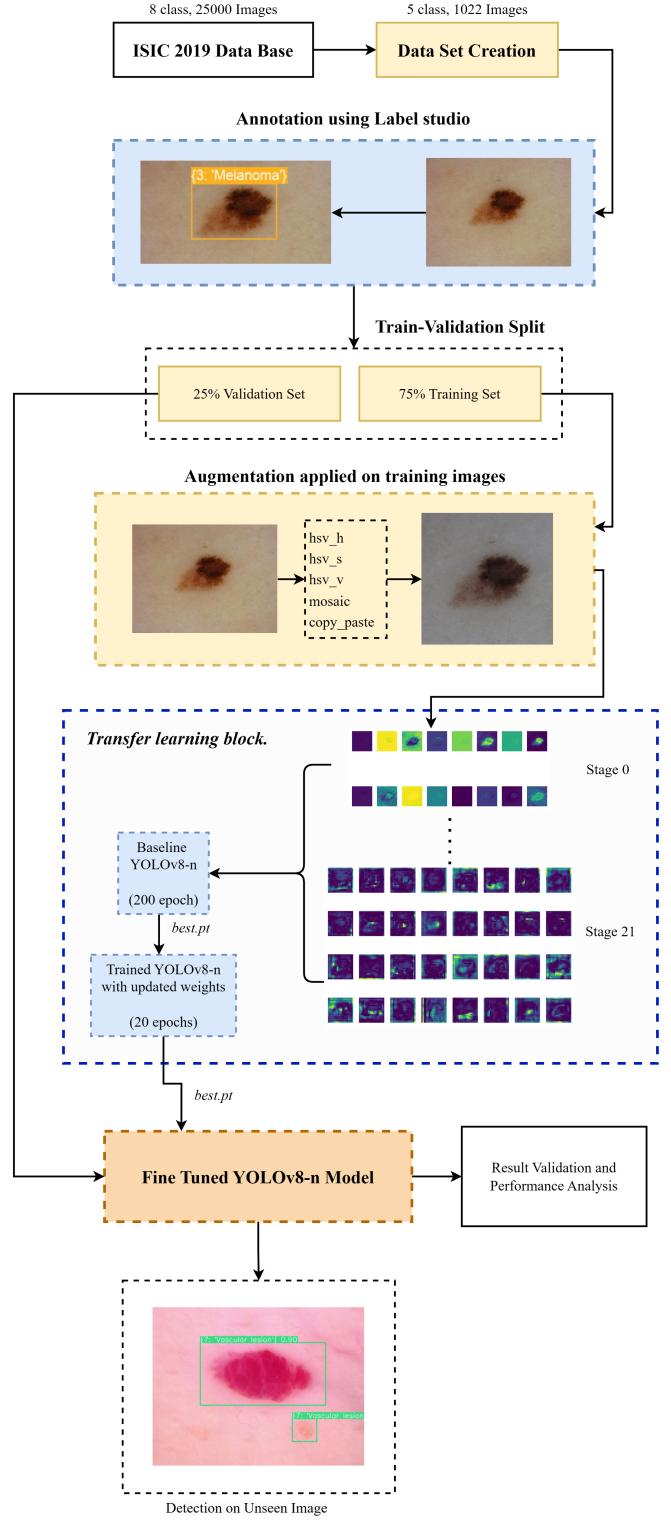


Fig. 1. Workflow diagram with visualization

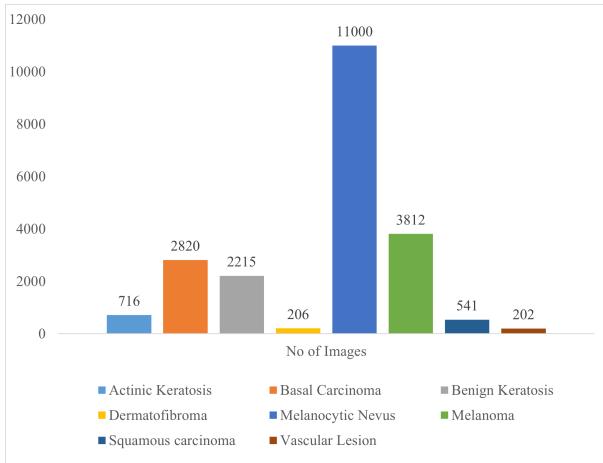


Fig. 2. Number of images per class in original ISIC 2019 training dataset.

III. METHODOLOGY

A. Dataset

The dataset used here is originally from the famous ISIC 2019 challenge [19]–[21]. The International Skin Imaging Collaboration (ISIC), is a collaborative effort from different parts of the world for better understanding skin with technologies. It is backed by the International Society for Digital Imaging of the Skin (ISDIS). Since 2016, ISIC has launched several skin cancer image datasets, which are publicly available. The 2019 dataset had more than 25000 images from eight different types of skin cancers and was collected from Kaggle [22]. However, it was heavily imbalanced, as a single cancer class had more than 11000 images, whereas another had only 200 images. To reduce the dimension of the dataset and imbalance handling, we took 1022 images from 5 different classes, averaging 200 images from each class. The classes were: actinic keratosis, benign keratosis, melanocytic nevus, melanoma and vascular lesion. After the images were selected, they had to be annotated to work with YOLO.

B. Annotation

There was no publicly available annotated ISIC 2019 dataset. So, we annotated the images ourselves with the supervision of a medical professional. An open-source tool called Label Studio was used for annotation task [23]. All of the 1022 images were annotated with bounding boxes around the cancerous regions. YOLO requires the labels to be stored separately as a text file for all the images. Label Studio performed the image and label storing after the annotations were completed. This bounding box information is considered the ground truth.

C. Dataset Splitting

The annotated dataset was then split into a 75:25 ratio, 75% data was used for training and 25% data was used for validation. To ensure the split is balanced, 50 images were taken from each class in the validation set. So, 772 images

are for training and the rest for validation. According to the official documentation of YOLO, it is recommended to have at least 1500 images per class and 10000 instances per class [24]. The limited data availability made this multi-class skin cancer detection very challenging.

D. YOLOv8 Object Detection Algorithm

YOLOv8 was introduced in 2023 and includes three primary components. These are classified as a backbone network, a feature extraction network, and a detection network [25]. The backbone model used in YOLOv8 is known as CSPDarknet53, and it is a modified version of the VGG-16 model that incorporates cross-stage partial architecture (CSP) [26]. This improves YOLOv8's accuracy and efficiency. The basic idea of CSPDarkNet53 is to divide each residual block into two sections: the main branch and the side branch. The main branch performs convolutional operations and feature extraction, whereas the side branch reduces the number of channels using 1×1 convolution [27]. YOLOv8 have five architectures, each with more parameter counts than the last one [28]. We opted for YOLOv8-n or the nano model with 3.2 Million parameters. The main reason for choosing this nano variant was the small footprint of the dataset used. Usually, higher parameter models require more data to fit properly and generate usable output.

E. Transfer Learning Setup

To achieve the transfer learning approach, the base YOLOv8-n was fine-tuned. Fine-tuning involves taking a pre-trained model and further training it on a specific dataset, often with a change in hyperparameters. YOLOv8-n is trained on the COCO dataset and has a larger pre-trained weight base. These weights are fine-tuned for our cancer dataset. It is trained for 200 epochs first. This creates a baseline model. The saved best model was then loaded and trained for another 20 epochs on the shuffled training dataset. This helped the model to be more robust and learn underlying variations. This was observed from the training metrics. The image augmentation was applied both times, ensuring the model learns complex variations.

F. Image Augmentation

YOLOv8 takes a batteries-included approach, which means there are a lot of hyper-parameters to be tuned without explicitly programming in Pytorch. One of the key features is the ability to augment the images. In our task, we focused on dataset augmentation. The augmentation parameters are given in Table 1.

G. Performance Metrics

There are several metrics (Intersection over Union (IoU), Mean Average Precision (mAP), F1 score etc.) available to understand the YOLO model's performance. However, the class-wise metric enables a deeper understanding of an object detection model. Similar to classification reports, class-wise metrics can show overall performance and how accurately each class of the dataset is being detected by the model. As

TABLE I
APPLIED IMAGE AUGMENTATION PARAMETERS

Augmentation	Description	value
hsv_h	Adjusts the hue of the image by a fraction of the colour wheel, introducing colour variability. Helps the model generalize across different lighting conditions.	0.5
hsv_s	Alters the saturation of the image by a fraction, affecting the intensity of colours. Useful for simulating different environmental conditions.	0.5
hsv_v	Modifies the value (brightness) of the image by a fraction, helping the model to perform well under various lighting conditions.	0.5
mosaic	Combines four training images into one, simulating different scene compositions and object interactions. Highly effective for complex scene understanding.	0.5
copy_paste	Copies objects from one image and pastes them onto another, useful for increasing object instances and learning object occlusion.	0.5

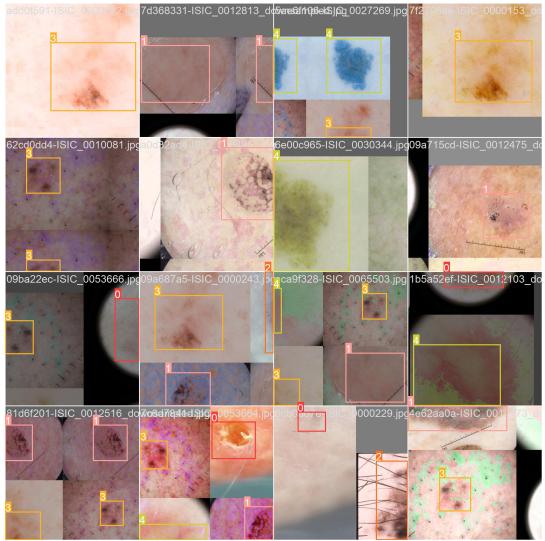


Fig. 3. Augmentation of a training batch

our dataset had five different skin cancer classes, using class-wise metrics was an ideal solution. This had four different performance information, which are: Precision (P), recall (R), mAP50 and mAP50-95.

Precision is simply the accuracy of the detected objects in the images. Higher precision means the model could correctly identify most of the objects. Recall, on the other hand, is crucial for multi-class classification. It shows the model's effectiveness in identifying all instances of a certain object in all the images. mAP50 is the mean average precision at the IoU threshold set to 0.5. This metric encompasses both precision and IoU, which is essential in understanding the accuracy of comparatively simpler detections. mAP50-95 is a similar metric, but the IoU threshold is varied from 0.5 to 0.95. This entices accuracy in both simple and complex detections. All these metrics were calculated for overall performance and

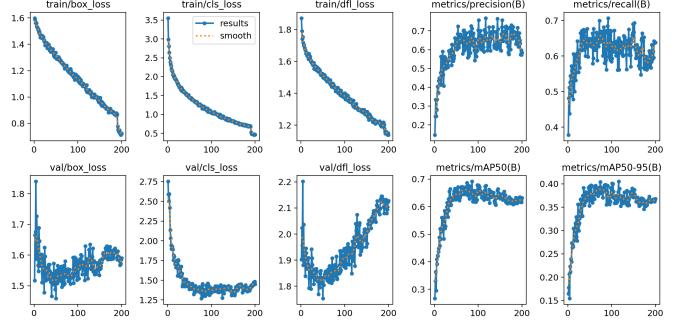


Fig. 4. Training YOLOv8-n for 200 epochs for baseline model creation

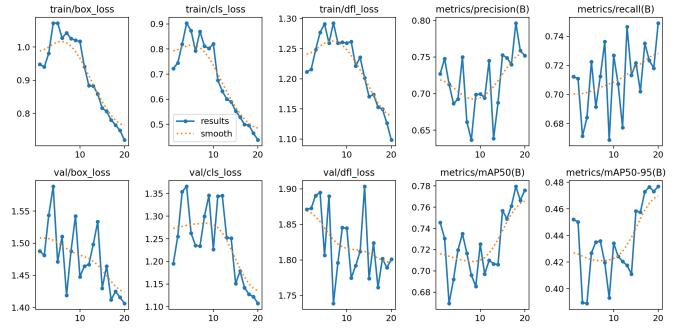


Fig. 5. Training the best baseline model for 20 epochs

each class detection for all the models.

IV. RESULT ANALYSIS

A. Training

Our fine-tuning approach saw a drastic improvement in training metrics. During the first training, the baseline YOLOv8-n had a mAP50 of 65%. Also, the precision and recall showed an unstable nature, reaching a maximum of 70%. After the first phase of training had been completed, the best-trained model was taken and loaded as a new model. This time the model ran for only 20 epochs and showed a promising result. The precision and recall showed a constant improvement. Training mAP50 increased to 80%, with an uprising trend. All the losses, along with the distribution focal loss (dfl_loss) were decreasing too.

B. Validation

The trained best model was then used for validation. During validation, all the classes were identified with very high precision and recall. The correlation matrix shows that there were very few misclassifications. Vascular lesions were detected with the most mAP50, other classes were very close as well. From the validation output, it was seen that the model was able to detect multiple regions and classes on a single image. Also, the number of detection bounding boxes was higher than the ground truths in some instances, which means the augmentation helped discriminate more regions from the

TABLE II
PERFORMANCE METRICS FROM VALIDATION SET

Class	P	R	mAP50	mAP50-95
Actinic keratosis	0.981	0.991	0.968	0.477
Benign Keratosis	0.989	0.992	0.973	0.512
Melanocytic Nevus	0.991	0.992	0.98	0.482
Melanoma	0.994	0.986	0.987	0.49
Vascular Lesion	1	1	1	0.62
All	0.991	0.992	0.982	0.516

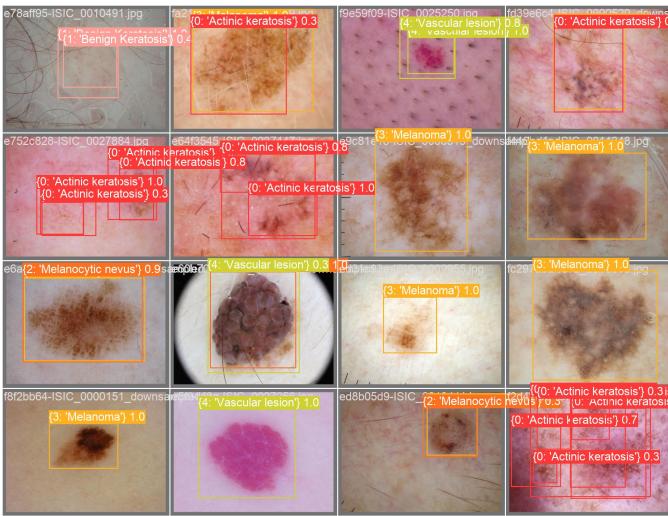


Fig. 6. Bbox on predictions of a validation batch image

affected area of the image. A similar result was found in completely unseen images as well.

From the training and validation, we observed that the model, due to augmentation, was able to perform very well. During training, the lower metrics could be due to the imbalance in picture size of different class images. Also, the instances being similar in several cases made it tough to train the model. Despite these limitations in training, in most pictures, the cancer region detection was accurate. In some images, several regions were identified, even if the ground truth was only one instance. This increase in bbox led to higher performance metrics in the validation set than in the training set. We also saw our approach was overfitting on the validation set due to the small validation size.

C. Detecting on Unseen Image

The final model was tested on several images from the original ISIC dataset, which were not annotated and weren't a part of the training or validation. In most cases of actinic keratosis, vascular lesions and benign keratosis were detected perfectly. Melanoma and melanocytic nevus saw a somewhat mixed result. This testing on unseen images shined on the trustworthiness of the approach.

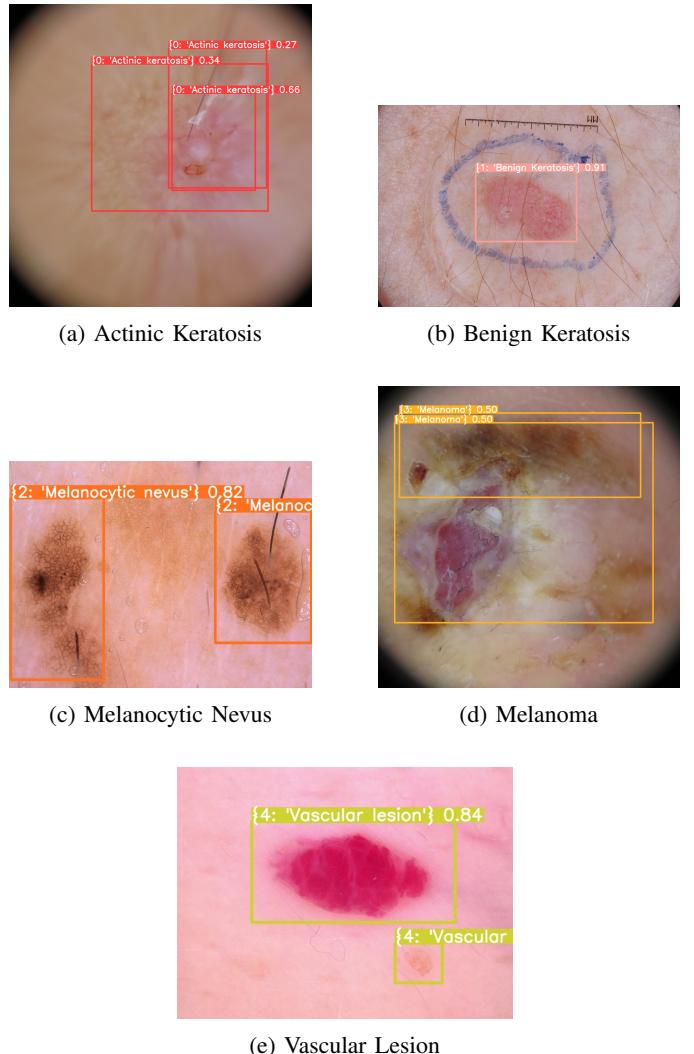


Fig. 7. Detection on unlabelled, out-of-set images.

V. CONCLUSION

This study is aimed at establishing a novel approach to skin cancer lesion detection from a small dataset. YOLOv8 is a state-of-the-art algorithm for object detection, but the number of annotated ground truths available here was very much smaller than typically suggested for training YOLOv8. This data shortage challenge was overcome by fine-tuning image augmentation parameters for synthetic training data generation. To achieve a more generalized trained object detector model, a similar method of transfer learning was opted for. Creating a baseline YOLOv8 and then repeating the training with the baseline model ensures that the final model is optimized with all the different variations of training images. This work is then concluded with testing and validating the model on unseen images, which shows praiseworthy performance in detection. Our approach can match the performance of several large models, trained on much larger datasets. However, there is still room for improvement, as we saw the

model hallucinating over similar images. Also, a more uniform dataset will prove to be more effective for the training of YOLOv8.

REFERENCES

- [1] H.-W. Huang, B. W.-Y. Hsu, C.-H. Lee, and V. S. Tseng, "Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers," *The Journal of dermatology*, vol. 48, no. 3, pp. 310–316, 2021.
- [2] B. S. Kim, M. Ahn, W.-W. Cho, G. Gao, J. Jang, and D.-W. Cho, "Engineering of diseased human skin equivalent using 3d cell printing for representing pathophysiological hallmarks of type 2 diabetes in vitro," *Biomaterials*, vol. 272, p. 120776, 2021.
- [3] V. W. Rebecca, R. Somasundaram, and M. Herlyn, "Pre-clinical modeling of cutaneous melanoma," *Nature communications*, vol. 11, no. 1, p. 2858, 2020.
- [4] S. Albahli, N. Nida, A. Irtaza, M. H. Yousaf, and M. T. Mahmood, "Melanoma lesion detection and segmentation using yolov4-darknet and active contour," *IEEE access*, vol. 8, pp. 198 403–198 414, 2020.
- [5] A. Magdy, H. Hussein, R. F. Abdel-Kader, and K. Abd El Salam, "Performance enhancement of skin cancer classification using computer vision," *IEEE Access*, 2023.
- [6] P. Ghosh, S. Azam, R. Quadir, A. Karim, F. Shamrat, S. K. Bhowmik, M. Jonkman, K. M. Hasib, and K. Ahmed, "Skinnet-16: A deep learning approach to identify benign and malignant skin lesions," *Frontiers in Oncology*, vol. 12, p. 931141, 2022.
- [7] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [8] S. Bechelli and J. Delhommele, "Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images," *Bioengineering*, vol. 9, no. 3, p. 97, 2022.
- [9] S. Maqsood and R. Damaševičius, "Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare," *Neural networks*, vol. 160, pp. 238–258, 2023.
- [10] R. Druon, Y. Yoshiyasu, A. Kanezaki, and A. Watt, "Visual object search by learning spatial context," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1279–1286, 2020.
- [11] R. Huang, J. Pedoeem, and C. Chen, "Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers," in *2018 IEEE international conference on big data (big data)*. IEEE, 2018, pp. 2503–2510.
- [12] N. Aishwarya, K. M. Prabhakaran, F. T. Debebe, M. S. S. A. Reddy, and P. Pranavee, "Skin cancer diagnosis with yolo deep neural network," *Procedia Computer Science*, vol. 220, pp. 651–658, 2023.
- [13] R. Tanna and T. Sharma, "Binary classification of melanoma skin cancer using svm and cnn," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*. IEEE, 2021, pp. 1–4.
- [14] A. Hekler, J. S. Utikal, A. H. Enk, A. Hauschild, M. Weichenthal, R. C. Maron, C. Berking, S. Haferkamp, J. Klode, D. Schadendorf *et al.*, "Superior skin cancer classification by the combination of human and artificial intelligence," *European Journal of Cancer*, vol. 120, pp. 114–121, 2019.
- [15] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets—a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.
- [16] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Computer methods and programs in biomedicine*, vol. 193, p. 105475, 2020.
- [17] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020.
- [18] S. K. Singh, V. Abolghasemi, and M. H. Amisi, "Fuzzy logic with deep learning for detection of skin cancer," *Applied Sciences*, vol. 13, no. 15, p. 8927, 2023.
- [19] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "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)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.
- [20] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermatoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.
- [21] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.
- [22] B. prasanna, "Isic - 2019," Dec 2022. [Online]. Available: <https://www.kaggle.com/datasets/bhanuprasanna/isic-2019>
- [23] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020–2022, open source software available from <https://github.com/heartexlabs/label-studio>. [Online]. Available: <https://github.com/heartexlabs/label-studio>
- [24] G. Jocher, "Tips for best training results," Jan 2024. [Online]. Available: https://docs.ultralytics.com/yolov5/tutorials/tips_for_best_training_results
- [25] Q. Zhao, B. Li, and T. Li, "Target detection algorithm based on improved yolo v3," *Laser & Optoelectronics Progress*, vol. 57, no. 12, p. 121502, 2020.
- [26] P. Wang, Y. Tang, F. Luo, L. Wang, C. Li, Q. Niu, and H. Li, "Weed25: A deep learning dataset for weed identification," *Frontiers in Plant Science*, vol. 13, p. 1053329, 2022.
- [27] Q. He, A. Xu, Z. Ye, W. Zhou, and T. Cai, "Object detection based on lightweight yolox for autonomous driving," *Sensors*, vol. 23, no. 17, p. 7596, 2023.
- [28] T. Najib, F. Muntasir, and W. A. W. Wasi, "Benchmark study on yolov8 variants in localized multiclass fault detection in pcbs," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*. IEEE, 2024, pp. 562–567.