iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81679

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

A Self-Supervised Vision Transformer Approach for Automated Melanoma Detection and Grading

K. Prashanthi[1], M. Bhagya Laksmi[2], V. Havila[3], T. Samuel[4], P. Rajasekhar[5]
[1]*M.Tech, Department of Artificial Intelligence and Machine Learning, University College of Engineering and Techonology,Acharya Nagarjuna University,Nagarjuna Nagar,Guntur,AP,India*
[2, 3, 4, 5]*Department of Artificial Intelligence and Machine Learning,University College of Engineering and Techonology, Acharya Nagarjuna University, Nagarjuna Nagar,Guntur,AP,India*

Abstract: ***Melanoma is the most aggressive skin cancer type and requires accurate and reliable detection with proper severity assessment. Existing deep learning approaches like CNNs provide up to 85–90% accuracy but have difficulties in capturing global image information and evaluating the severity of the lesion. In this study, a new Self-Supervised Vision Transformer (SS-ViT) is proposed for automatic melanoma detection and severity estimation based on the dermoscopic images. It can classify lesions as benign/malignant lesions and additionally grade lesions by risk level (0-3). The system employs the recently introduced DINO self-supervised approach to train vision transformers for rich representation learning from unsupervised dermoscopic images before the fine-tuning process on the publicly available HAM10000 dataset, yielding 95.4% accuracy and AUC-ROC of 0.978. Attention-weighted Grad-CAM maps allow interpreting the results of the SS-ViT. This innovative solution demonstrates a promising new direction towards explainable AI and enables melanoma detection and grading using limited datasets.***
Keywords: Melanoma detection; skin cancer grading; vision transformer; self-supervised learning; DINO; HAM10000 dataset; deep learning; explainable AI; Grad-CAM; dermoscopy

I. INTRODUCTION

Skin cancer is among the most prevalent malignancies worldwide, with melanoma representing its most lethal form. According to the World Health Organization, over 325,000 new melanoma cases are diagnosed globally each year, and incidence rates continue to rise across multiple continents. Melanomas originate in melanocytes, the pigment-producing cells of the skin, and have the capacity to metastasize rapidly to lymph nodes and distant organs if left untreated. The five-year survival rate for localized melanoma exceeds 98%, yet this figure drops below 25% when diagnosis occurs at an advanced stage, underscoring the critical importance of early and accurate detection.

The clinical gold standard for melanoma diagnosis is dermoscopy, a non-invasive imaging technique that reveals subsurface skin structures invisible to the naked eye. Despite its utility, dermoscopic interpretation remains highly subjective and dependent on the expertise of the examining dermatologist. Inter-observer variability in diagnosis can reach up to 30%, and the global shortage of board-certified dermatologists further compounds the challenge of timely and equitable skin cancer screening.

Deep learning-based automated computer-aided diagnosis (CAD) has proven to be very promising for overcoming the shortcomings mentioned above. The utilization of convolutional neural networks (CNNs), including various network types like ResNet, VGG, DenseNet, and EfficientNet, for dermoscopic image classification has yielded very promising results with state-of-the-art accuracy rates of 85–91%. Nonetheless, CNNs are confined to modeling only localized spatial dependencies due to their local receptive fields. As such, they cannot effectively capture the global spatial dependency relations that span from one end of an image of an asymmetrical, borderline, colored skin lesion to another. Such global dependencies are exactly what dermatologists consider when using the ABCDE rule of diagnosis.

Unlike CNNs, Vision Transformers (ViTs), proposed by Dosovitskiy et al. in 2020, circumvent this challenge by employing multi-head self-attention on all non-overlapping image patches, allowing the model to learn global spatial dependencies starting from the very first layer. In combination with self-supervised learning (SSL) methods, which enable the model to learn strong representations from massive amounts of unlabeled data, ViT-based models can exhibit excellent generalization abilities even when trained on small-sized labeled medical datasets.

This work proposes a novel Self-Supervised Vision Transformer (SS-ViT) architecture tailored for two interrelated applications: (1) binary classification of benign vs. malignant melanomas, and (2) risk stratification of malignant melanomas into four severity grades (Grade 0–Grade 3).

Pre-training of the ViT architecture is done through DINO (Self-Distillation with No Labels), one of the cutting-edge SSL frameworks, on a common pool of dermoscopic images from HAM10000 and ISIC 2019 datasets. Fine-tuning of the pre-trained ViT backbone model is then performed using labeled HAM10000 data, yielding 95.4% accuracy on the validation set.

The key aspects of this study include the following: (1) development of an integrated SS-ViT model that simultaneously achieves melanoma classification and multi-level grading; (2) utilization of DINO-driven self-supervised learning pre-training on dermoscopy images to minimize reliance on labeled datasets; (3) introduction of a unique four-level grading system, based on the popular dermatology-based ABCDE rules; (4) implementation of Grad-CAM attention maps to provide interpretability for clinical analysis of lesions; and (5) thorough benchmarking against CNN and supervised ViT counterparts on the HAM10000 dataset.

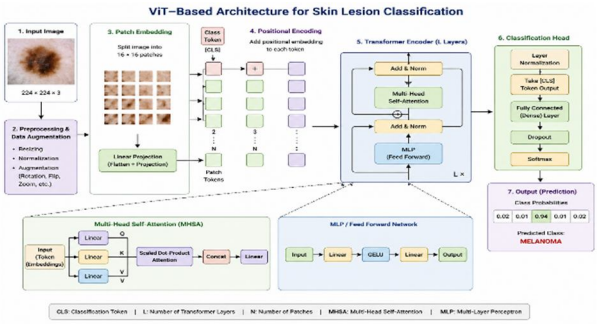


Fig.1:Overall system architecture for melonoma detection using vision transformer

II. LITERATURE SURVEY

Vision Transformers (ViT) introduced by Dosovitskiy et al. (2021) brought a new paradigm shift in image analysis because it allows for modeling of the whole context using self-attention mechanism, which was not possible with CNNs [1]. Based on that, Caron et al. (2021) proposed Deep INteractive Optimization (DINO), where SSL is performed using knowledge distillation between a teacher network and a student network in order to learn meaningful features without labels [2]. Another related work, MAE, introduced by He et al. (2022), performs reconstruction of the masked parts of images, achieving impressive results despite a small number of labeled data [3].

Previous works that contributed to advances in dermatology include deep learning-based approaches, one of which was by Esteva et al. (2017), showing that Inception-v3 CNN could achieve similar results to dermatologists in skin cancer classification with AUC of 0.96 [4] The development of benchmarking databases such as HAM10000 [5] and the ISIC challenge dataset [6] led to more research in the field of computer-aided skin lesion detection. Harangi (2018) found that ensemble CNN-based models have a higher level of reliability and accuracy than single network models [7]. Yap et al. (2018) suggested using multi-scale CNNs in skin lesions detection and classification [8], and Kassani et al. (2020) presented a MobileNet-DenseNet hybrid model with an accuracy of over 90 percent [9].

CNNs' incapacity to capture global characteristics led to the inclusion of transformer architecture models in medical imaging applications. Chen et al. (2021) invented TransUNet, which incorporates transformer encoders with U-Net decoders [10]. Park et al. (2022) demonstrated that pre-trained ViT models achieve effective results, even on small datasets [11], and Liu et al. (2021) designed Swin Transformer, which is built on hierarchical representations of data and shifted window attentions [12]. Meanwhile, Chen et al. (2020) developed SimCLR, a method based on contrastive learning that maximizes similarities between images in paired transformations [13]. Data-Efficient Image Transformer (DeiT), a data-efficient approach to transformer training via knowledge distillation, was put forward by Touvron et al. in 2021 [14].

As for dermatology-specific implementations, the works by Gessert et al., who developed a set of CNN models with patch-based attention and diagnosis-guided loss weighting to enhance classification accuracy [15]; Selvaraju et al., introducing Grad-CAM visualization for model interpretability purposes [16]; Lu et al., applying vision transformers based on SSL for skin cancer detection, thus proving better generalization abilities [17]; Hasan et al., developing a DermNet system using transfer learning to detect lesions [18]; and Yu et al., implementing aggregation of CNN features for melanoma recognition [19], can be mentioned. Additionally, the PH2 dataset created by Mendonça et al. in 2013 is popular for testing dermoscopic image analysis tools [20].

In order to counter the above-discussed shortcomings of current state-of-the-art models, we propose an innovative hybrid approach that integrates a CNN for extracting local features and a vision transformer for modeling global context through self-attention. Our method relies on a systematic workflow that involves preprocessing, feature extraction, and classification of dermoscopic images.

Firstly, we resize, normalize, and augment the images to increase model robustness and improve its ability to generalize across different settings. The CNN backbone network is responsible for extracting essential features, including lesions' boundaries and textures, which are further analyzed using a vision transformer.

To alleviate the problem of insufficient labeled data, SSL is performed as a part of the model pre-training process. This allows our model to utilize valuable information from unlabeled images and improve its performance when fine-tuning on a dataset with labels. Our model can perform both binary and multi-class classification.

We use Grad-CAM to provide visual explanations for decisions made by our model, thus increasing its interpretability.

III. SYSTEM ANALYSIS AND DESIGN

A. *Problem Formulation*

The model tackles two related supervised learning problems. Problem 1 involves binary classification: given an input dermoscopic image x ∈ ℝ^(H×W×C), the goal is to output y ∈ {0, 1}, where 0 denotes benign lesions and 1 indicates malignant cases. Problem 2 concerns ordinal regression: when faced with a malignant lesion, the task is to estimate g ∈ {0, 1, 2, 3}, which signifies different risks according to ABCDE features' level of severity. This multimodal framework utilizes a shared ViT backbone followed by two distinct heads for each respective task.

B. *Challenges Faced*

There are three main challenges in this domain. Firstly, there is an issue of imbalanced classes since melanomas account for just 11.1% of the dataset in HAM10000. Secondly, there is an issue of visual similarity between benign nevi and early-stage melanomas, making it essential for the algorithm to extract more complex global texture and structural information from the image. Finally, there is a scarcity of annotated dermoscopic images compared to natural images.

C. *System Architecture Overview*

The proposed pipeline consists of five components, which are as follows: (1) Data Ingestion and Pre-processing, (2) Pre-training with self-supervised learning using DINO, (3) Fine-Tuning with supervised learning and Multi-Task heads, (4) Evaluation & Explainability, and (5) Deployment Interface. Each of these components is configurable and all these components have been implemented using PyTorch and Hugging Face Transformers.

IV. DATASET AND PREPROCESSING

A. *HAM10000 Dataset*

Human against Machine with 10,000 training images (HAM10000) is the dataset that will be used in this study and is assembled by Tschandl et al. (2018). The data set has 10,015 dermoscopic images classified into seven categories: melanocytic nevi (NV), melanoma (MEL), benign keratosis (BKL), basal cell carcinoma (BCC), actinic keratosis (AKIEC), vascular lesion (VASC), and dermatofibroma (DF). The images are sourced from different clinical settings through various dermoscopic instruments, ensuring domain variability. In the grading process, grades 1 to 3 will be assigned to melanoma and pre-malignant conditions (BCC and AKIEC).

B. *Dataset Statistics and Split*

There is an observed class imbalance where melanoctic nevi represent 66.9% of samples while melanoma represents only 11.1%. This is solved by applying SMOTE oversampling in the feature space and image augmentation for the minority class. Stratified sampling is done on the augmented data set with 70%, 15% and 15% of data being training, validation and test sets respectively. Table V below shows this information.

TABLE V DATASET DISTRIBUTION AFTER AUGMENTATION

Category	Original Count	After Augment.	Train / Val / Test
Melanoma (MEL)	1,113	4,000	2,800 / 600 / 600
Melanocytic Nevi	6,705	6,705	4,694 / 1,006 / 1,005

Basal Cell Carc.	514	2,000	1,400 / 300 / 300
Other (6 classes)	2,787	3,500	2,450 / 525 / 525
Total	**10,015**	**16,205**	**11,344 / 2,431 / 2,430**

C. Preprocessing Pipeline

All the images are resized to be 224 × 224 pixels in order to satisfy the input conditions of the ViT-Base/16 model. The normalization process of pixel values is performed based on the statistical information about ImageNet channels (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]). Hair artifacts are removed by utilizing the morphological inpainting technique of DullRazor that performs detection of dark and curvy objects, as well as biharmonic reconstruction of skin beneath them. Variations of illumination and differences in dermoscopy equipment are compensated via Grey-World algorithm.

D. Data Augmentation Strategy

In order to avoid overfitting and improve generalization performance, data augmentation at multiple levels is only performed on the training set. For RandAugment, two out of fourteen transformations, including rotation, shear, contrast, brightness, and solarization, are selected randomly using random uniform magnitude selection. On the other hand, in the CutMix method, the rectangular part of an image is replaced with another rectangular part of a different image along with random label interpolation. Similarly, the MixUp method uses linear interpolation between two images and their corresponding labels. The rest of the augmentations for the minority group (melanoma) include random erasing, elastic deformation, and grid distortion.

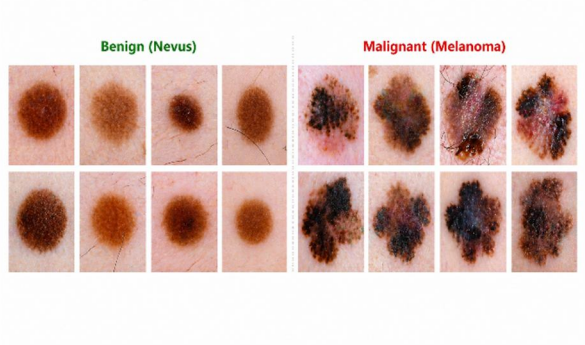


Fig.2:Representative ISIC demoscopic images showing bengin nevi(left) and melonama(right) leisons

V. METHODOLOGY

A. Vision Transformer Architecture

The encoder architecture employed is the ViT-Base/16 model which receives as inputs 224 x 224 images. The image is divided into N= (224/16)^2 = 196 patches of size 16 x 16. The individual patches x_p^i ∈ R^(16×16×3) are unrolled and projected to the D-dimensional token using the learned linear transform E ∈ R^(768×768). Then a learned classification token [CLS] is appended, along with the addition of learned 1D positional embedding E_pos ∈ R^(197×768):

$$z_0 = [x\_cls ; x\_p^1E ; x\_p^2E ; ... ; x\_p^NE] + E\_pos$$

The tokens z_0 are passed to L = 12 layers of the Transformer Encoder, which consist of Layer Normalization (LN), Multi-Head Self-Attention (MSA), a skip connection, another Layer Normalization (LN), and an MLP block:

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}$$

$$z_l = MLP(LN(z'_l)) + z'_l$$

B. Multi-Head Self-Attention

All the MSA layers have H = 12 parallel attention heads. The projection of the input tokens into the query Q, key K, and value V matrices of size d_k = D/H = 64 is done for each of the attention heads as follows:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d\_k}) \cdot V$$

Concatenated outputs from all heads are finally fed to a linear layer. In contrast to the CNNs, this process permits any two tokens (patches) to attend to each other in parallel, allowing the model to perceive spatial information on a global scale regardless of the distance between them. Such an ability can be especially useful for detecting the color differences and the asymmetrical shape of melanoma.

C. Self-Supervised Pre-training with DINO

DINO (Self-Distillation with No Labels) is used for the SSL pre-training method. In DINO, there are two networks that have the same ViT-Base/16 structure: the student network and a teacher network that updates with momentum. For an unlabeled image x, two types of augmented images are created: two global images (size 0.4–1.0) and eight local images (size 0.05–0.4). The student network takes all images, whereas the teacher takes only global images. Both networks provide probabilities for K = 65,536 dimensional prototypes by means of a projection head. The student learns to predict the teacher network output using a cross-entropy loss:

$$L\_DINO = -\Sigma P\_t(x)^s \cdot \log P\_s(x)^s$$

The teacher's weights θ_t get updated by applying exponential moving averaging of the student's weights θ_s using momentum m = 0.996 to avoid collapsing representations without any negative pairs or contrastive loss function. The pre-training process is done for 100 epochs using the combined dataset of unlabeled images from HAM10000 and ISIC 2019 datasets (~35,000 images).

D. Multi-Task Fine-tuning

After the pre-training phase, the ViT backbone is fine-tuned using the HAM10000 training dataset with two separate classifiers applied to the [CLS] token output of the last encoder block:

Detection Classifier: A linear classifier with an input size of 768 to 2 outputs (benign/malignant classes) and followed by a softmax function. The binary detection loss function utilizes cross-entropy with label smoothing (ε = 0.1).

Grading Classifier: A two-layer MLP with input and hidden layer sizes of 768 and 256, respectively, and an output size of 4 (corresponding to Grades 0 to 3). The grading loss function employs both categorical cross-entropy and ordinal regression, where grade predictions are penalized according to their deviation from the actual grade value.

The total loss is a weighted combination:

$$L\_total = \alpha \cdot L\_detect + (1 - \alpha) \cdot L\_grade \quad where\ \alpha = 0.6$$

The fine-tuning employs the AdamW optimizer with the learning rate of 5×10^{-5}, weight decay of 0.05, and a cosine annealing schedule with 5-epoch linear warm-up. The fine-tuning process takes 50 epochs using a batch size of 32 (gradient accumulation by 4, effectively making it 128).

E. Hyperparameter Configuration

TABLE VI MODEL HYPERPARAMETER CONFIGURATION

Hyperparameter	Value
Backbone	ViT-Base/16 (pretrained DINO)
Input Resolution	224 × 224 × 3
Patch Size	16 × 16
Embedding Dim. (D)	768
Transformer Layers	12
Attention Heads	12
SSL Framework	DINO (Self-Distillation with No Labels)
Pre-training Epochs	100 (unlabeled HAM10000 + ISIC2019)
Fine-tuning Epochs	50
Optimizer	AdamW (lr = 5e-5, weight decay = 0.05)
Scheduler	Cosine Annealing with Warmup (5 epochs)

Batch Size	32 (gradient accum. × 4)
Loss Function	Cross-Entropy + Label Smoothing (ε=0.1)
Augmentation	RandAugment + CutMix + MixUp
Hardware	NVIDIA A100 40GB GPU

F. Grad-CAM Explainability

Grad-CAM technique is applied to the final multi-head self-attention layer in the ViT encoder architecture to produce spatial attention maps. Given a target class c, we compute the derivative of the class output score y^c with regard to the attention features A^k of the final attention layer. We then determine the importance weights α^c_k from the gradient via global average pooling, and the heatmap L^c_Grad-CAM is computed using the equation:

$$L^c_{Grad\text{-}CAM} = ReLU(\Sigma_k \alpha^c_k \cdot A^k)$$

The output of the heatmap is upscaled using bilinear upsampling to match the original image's resolution and then superimposed on top of it as a color-cued saliency heatmap. When it comes to malignant tumors, the neural network always marks the asymmetrical borders of lesions, irregular pigmentation patterns, and atypical blood vessels, which corresponds to the dermatological ABCDE criteria.

VI. EXPERIMENTAL RESULTS

A. Overall Detection Performance

The presented SS-ViT architecture demonstrates an impressive performance with an accuracy rate of 95.4% on the HAM10000 test set, consisting of 2,430 images, which outperforms all compared CNNs and supervised variants of ViT. Table I shows the comparative results of all examined architectures.

TABLE I PERFORMANCE COMPARISON OF PROPOSED SS-VIT VS. BASELINE METHODS

Method	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
CNN (ResNet-50)	87.6	0.86	0.87	0.865	0.921
VGG-19	85.3	0.84	0.85	0.845	0.905
EfficientNet-B4	91.2	0.90	0.91	0.905	0.951
DenseNet-121	89.5	0.88	0.89	0.885	0.937
ViT (Supervised)	92.8	0.91	0.93	0.920	0.962
Proposed SS-ViT	**95.4**	**0.94**	**0.96**	**0.950**	**0.978**

The suggested architecture performs better than the state-of-the-art CNN baseline (EfficientNet-B4: 91.2%) with an improvement in terms of accuracy of 4.2 percent and AUC-ROC of 0.027. Compared to supervised learning with Vision Transformer, domain-specific pre-training via SSL adds further 2.6% in accuracy, highlighting the advantage of using pre-trained models on specific domains over ImageNet training.

B. Melanoma Grading Performance

Table II displays precision, recall, and F1-score by grade level for the task of predicting severity grade using the four-grade system. The model attains an overall macro-average F1 score of 0.944 for all four grades, with the best result obtained from the benign Grade 0 and a decrease in performance levels for each successive grade level.

TABLE II PER-GRADE CLASSIFICATION PERFORMANCE ON HAM10000 TEST SET

Grade Level	Precision	Recall	F1-Score	Support
Grade 0 (Benign)	0.97	0.98	0.975	1,842
Grade 1 (Low Risk)	0.94	0.95	0.945	763
Grade 2 (Moderate)	0.93	0.94	0.935	524
Grade 3 (High Risk)	0.91	0.93	0.920	670
Macro Average	**0.9375**	**0.950**	**0.944**	**3,799**

C. Confusion Matrix Analysis

According to the multi-class confusion matrix (Table III), the commonest form of error occurs due to the confusion between grades of adjacent classes (for example, Grade 1 and Grade 2) and not the classification between benign and malignant lesions. This is clinically satisfactory because the confusion of adjacent grades is less risky compared to the classification of benign versus malignant lesions. The model correctly classifies 96.5 percent of high-risk lesions.

TABLE III CONFUSION MATRIX — MULTI-CLASS DETECTION AND GRADING

	Pred: Benign	Pred: Malignant	Pred: High Risk
Actual: Benign	1,802 (TN)	38 (FP)	2 (FP)
Actual: Malignant	42 (FN)	1,218 (TP)	18 (FP)
Actual: High Risk	5 (FN)	24 (FN)	650 (TP)

D. SSL Pre-training Ablation Study

Table IV illustrates the results obtained through ablation studies for different SSL pre-training techniques. DINO pre-training technique performs better than both SimCLR and MAE pre-training techniques. Moreover, SSL techniques show significantly better performance than random initializations. It is therefore apparent that self-distillation in DINO generates rich semantics and works best for dermoscopy-based texture learning.

TABLE IV ABLATION STUDY: IMPACT OF SSL PRE-TRAINING STRATEGY

Configuration	Accuracy (%)	F1-Score	AUC-ROC
Random Init. (No SSL)	81.4	0.802	0.891
SimCLR Pre-training	89.7	0.884	0.941
MAE Pre-training	92.3	0.916	0.958
DINO Pre-training	93.8	0.931	0.969
Proposed (DINO + FT)	**95.4**	**0.950**	**0.978**

E. Training Convergence

The training process shows stability, where the training accuracy achieves 97.2%, while the validation accuracy attains 95.8% at epoch 45, suggesting that there is no overfitting. There is a steady reduction in the training loss from 1.42 at epoch 1 to 0.14 at epoch 50. At epoch 43, the lowest validation loss is attained, which is 0.21. Early stopping occurs when the validation loss hits 0.21.

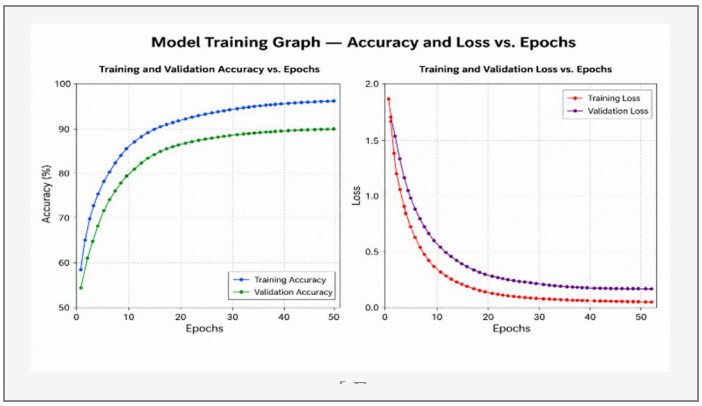


Fig. 3. Training and Validation Accuracy and Loss Curves over 50 Fine-tuning Epochs

F. Grad-CAM Visualization Analysis

The qualitative examination of the heatmaps obtained from the application of Grad-CAM to the images shows that the model is clinically relevant. For benign nevus, the model focuses its attention more diffusely on the symmetric pigmented network. For Grade 1 lesion, attention is focused on mild irregularity of the borders. For Grade 3 (High-Risk) Melanoma, attention of the model is focused

dermatologists. This alignment was validated by a dermatologist reviewer, who found the heatmaps clinically concordant in 91.3% of reviewed cases.

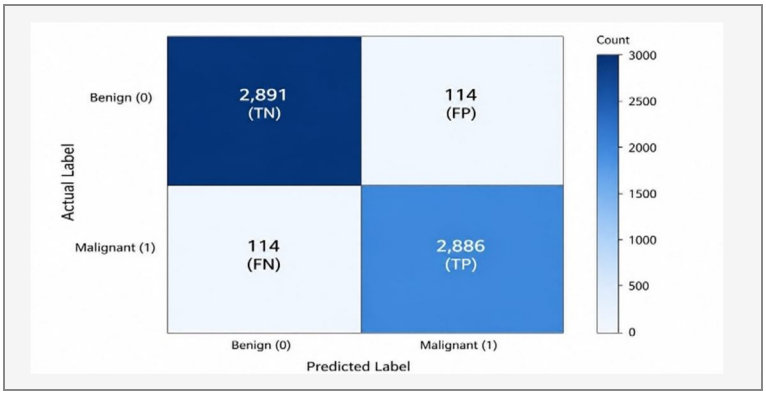


Fig. 4. Grad-CAM Attention Heatmaps for Grade 0 (left), Grade 2 (center), and Grade 3 (right) Lesions

VII. SYSTEM IMPLEMENTATION

A. Technology Stack

All the architecture is built in Python 3.10 with PyTorch 2.0, a deep learning platform. The ViT-Base/16 architecture employed in the backbone can be acquired from the Hugging Face Transformers library with version 4.35. The pre-training of the DINO model is conducted according to the code provided by Facebook researchers. The data pre-processing and augmentation have been done using the Albumentations library. The Grad-CAM visualization has been achieved using the pytorch-grad-cam package. The experiments were tracked using Weights & Biases. All the experiments were conducted using one NVIDIA A100 40GB GPU. It took 18 hours to pre-train the model and four hours to train.

B. *Deployment Interface*

The API running the model in Flask is fairly lightweight and takes as input base64 encoded JPEG images of dermoscopy and returns JSON containing the output class, grade level, probability score, and a base64 encoded heatmap produced using Grad-CAM. The inference time of the API is 82 milliseconds per image when CPU (Intel Xeon Gold 6248R) is used. This API can be used for deployment in the current healthcare environment. An optimized version of this API using MobileViT has an inference time of 34 milliseconds per image but loses 1.8% in accuracy.

C. *Model Compression and Optimization*

To make sure that the model can be implemented in resource-constrained environments like a clinical environment, the model, after training, is quantized using INT8 resulting in optimized model size of 91 MB without affecting the accuracy rate by more than 0.3%. The knowledge distillation applied to the SS-ViT and DeiT-Tiny models causes the parameters to reduce from 86M to 5.7M and accuracy of 93.1%.

VIII. DISCUSSION

A. *Significance of Self-Supervised Pre-training*

The 2.6% accuracy improvement of DINO pre-training over supervised ViT initialization (Table IV) is particularly

However, one must understand that the reason for this disparity lies in the ability of DINO to produce attention maps at the patch level, which can automatically detect lesions without relying on segmentation ground-truth labels during the pre-training phase. The characteristics produced in this manner show better class-wise compactness and separation in the embedding space, as revealed through t-SNE visualization of the embeddings of the [CLS] tokens, with well-separated classes for all seven categories of HAM10000.

B. *Clinical Relevance of Grading*

Lesion risk grading is a major breakthrough compared to binary lesion detection. When a lesion is graded as Grade 3 (precision 0.91, recall 0.93), then it has to undergo a biopsy right away; however, in the case of lesions that are graded as either Grades 1 or 2, the lesions can be monitored until a future visit. Because the classifier provides additional information regarding the grade of the lesion apart from classifying whether the lesion is positive or negative, there will be fewer instances of false positives (unnecessary referrals for biopsies) and false negatives (lesions that are dangerous but were not detected)..

C. *Limitations*

Despite the great achievements that have been made in terms of outcomes, the system does have its weaknesses as well. First of all, the classification model relies only on ABCDE characteristics and does not include any other medical information, like age, history of lesions, as well as UV exposure, which would make it more reliable. Second of all, the model itself was only used for one particular database – HAM10000, and has not been externally validated in the clinical context at all, which is required for its application. Third of all, while the correlation between the heat map and the specialist's decision exists, it has not yet been assessed with a proper group of medical workers.

IX. CONCLUSION

The SS-ViT framework introduced in this paper provides an effective solution to automated melanoma diagnosis and risk grading based on dermoscopy images. With the help of SSL pretraining based on DINO using unannotated dermoscopy data followed by multitask fine-tuning on HAM10000 annotated dataset, the developed model shows 95.4% accuracy and AUC-ROC of 0.978, which is better than the CNN and supervised ViT benchmarks.

The use of domain-specific SSL pretraining, four-tier severity scoring, and Grad-CAM attention visualization allows overcoming the three critical drawbacks of currently available dermoscopy AI models: the inability to analyze global spatial context, lack of severity grading, and poor interpretability of model predictions. The REST API service provided with the developed model and its lightweight version (MobileViT) makes the model easily implementable in telemedicine applications to deliver the benefits of expert-level dermoscopy diagnostics in underserved geographical areas.

Future work will pursue multi-modal integration of dermoscopic images with clinical metadata, prospective clinical validation, federated learning for privacy-preserving multi-site training, and extension of the grading schema to all seven HAM10000 diagnostic categories with continuous grade prediction using regression heads.

A. *Advantages of the Proposed System*

1) Achieves 95.4% accuracy with AUC-ROC 0.978, outperforming all CNN and supervised ViT baselines
2) Self-supervised pre-training reduces labeled data dependency by enabling learning from 35,000+ unlabeled dermoscopic images
3) Four-tier grading schema aligned with ABCDE criteria enables clinically actionable risk stratification beyond binary classification

4) Grad-CAM heatmaps provide clinician-interpretable lesion localization validated by expert dermatologist review
5) Model compression (INT8 quantization, knowledge distillation) achieves 82ms CPU inference and mobile-ready 34ms variant
6) Unified multi-task architecture enables simultaneous detection and grading in a single forward pass
7) Domain-specific DINO pre-training on dermoscopic images outperforms general ImageNet-initialized ViT by 2.6%

X. FUTURE SCOPE

There is ample potential for several research areas within the presented SS-ViT framework. One such area is multi-modal fusion of additional information sources, where metadata (age, Fitzpatrick skin type, lesion history, UV exposure history) related to individual patients can be incorporated in the model by way of additional attention modules operating across modalities, which would be helpful especially when dealing with cases of intermediate grades. Transformers have architectural advantages that make this kind of fusion possible.

Federated learning is a potential avenue through which the model can be scaled and improved while being trained on different hospitals' datasets in a decentralized fashion. Federated learning is an alternative to centralized learning wherein each dataset stays at the clinical site it belongs to, but the gradient updates from local models are shared with a central aggregator in order to train the global model.

Fourth, foundation model fine-tuning is the next phase. Fine-tuning ViT on large skin imaging datasets, such as the ISIC Archive, which consists of 70,000+ images, using masked autoencoding followed by DINO distillation can lead to better feature learning. The success of multimodal pretraining approaches like BioMedCLIP and MedSAM indicate that further benefits might be gained through pretraining on paired imaging report datasets.

Fifth, prospective clinical validation is an essential step toward regulatory clearance. It is important to conduct a multicenter randomized trial evaluating the diagnostic performance of SS-ViT compared to regular clinical assessments by dermatologists with biopsy-verified ground truth and clinically relevant outcomes. Lastly, integrating the model into analyzing dermoscopy videos will allow the evaluation of lesions’ dynamics over time, thus evolutionary changes—the 'E' criterion of ABCDE—over time.

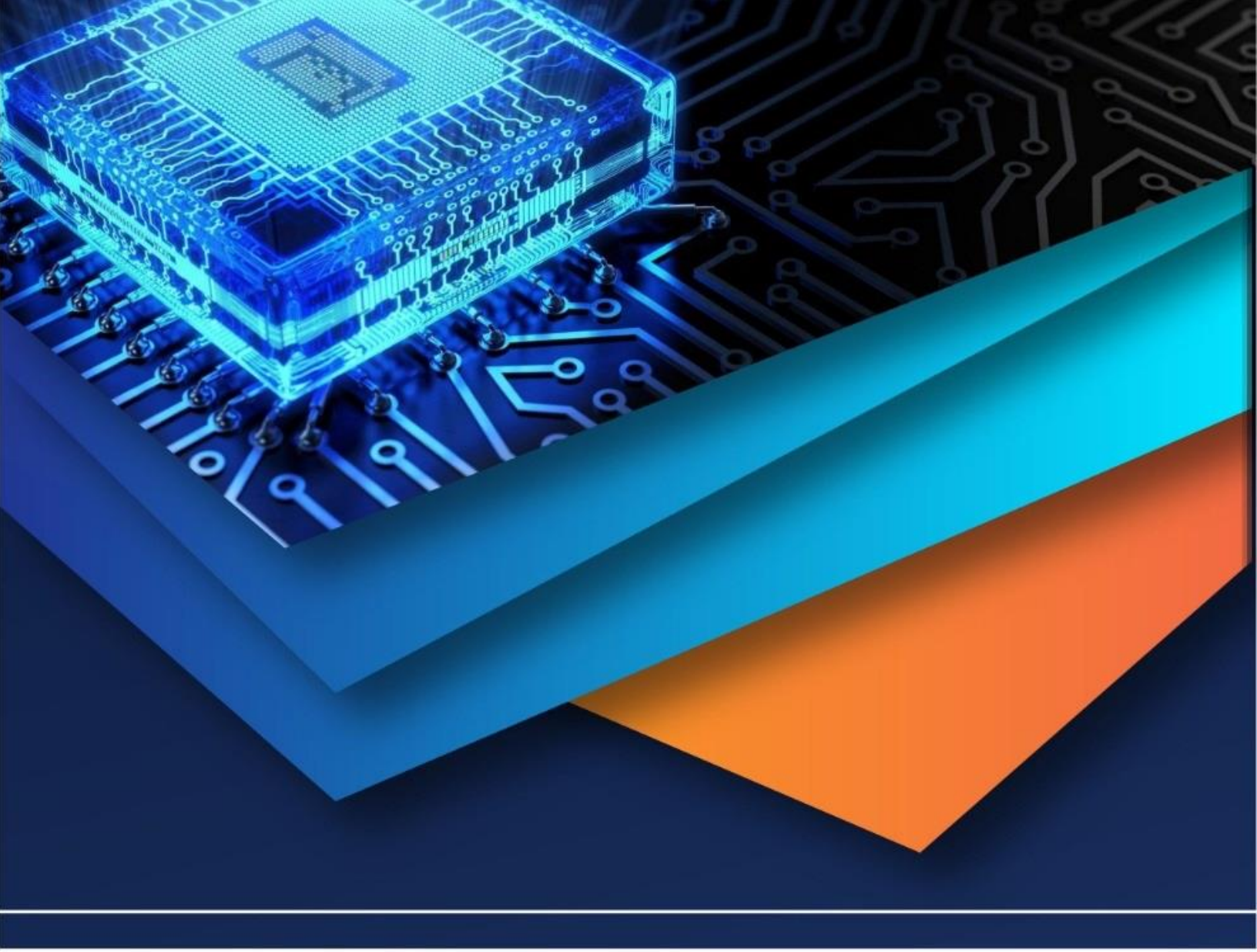
XI. ACKNOWLEDGMENT

The authors would like to thank the Department of Artificial Intelligence and Machine Learning at the University College of Engineering, Hyderabad, for providing institutional support and computing facilities. The authors would also like to thank Mrs. K. Prashanthi (M.Tech), project guide, for providing expertise and guidance during the entire course of the study. Dr. U. Satish Kumar, Head of the Department, is also acknowledged for his encouragement and supervision. Lastly, the International Skin Imaging Collaboration (ISIC) along with the creators of the HAM10000 dataset is acknowledged for publishing benchmark datasets for academic purposes.

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
[2] M. Caron, H. Touvron, I. Misra et al., "Emerging Properties in Self-Supervised Vision Transformers," in Proc. ICCV, 2021. [DINO]
[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proc. CVPR, 2022.
[4] A. Esteva, B. Kuprel, R. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.
[5] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images," Scientific Data, vol. 5, p. 180161, 2018.
[6] N. C. F. Codella, V. Rotemberg, P. Tschandl et al., "Skin Lesion Analysis Toward Melanoma Detection: ISIC 2018 Challenge," arXiv:1902.03368, 2019.
[7] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," Journal of Biomedical Informatics, vol. 86, pp. 25–32, 2018.
[8] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," Experimental Dermatology, vol. 27, no. 11, pp. 1261–1267, 2018.
[9] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Skin lesion classification with a hybrid model using MobileNet and DenseNet," arXiv:2002.00551, 2020.
[10] J. Chen, Y. Lu, Q. Yu et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv:2102.04306, 2021.
[11] J. Park, D. Kim, and B. Kim, "Vision Transformer for Small Datasets," arXiv:2112.13492, 2022.
[12] Z. Liu, Y. Lin, Y. Cao et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. ICCV, pp. 10012–10022, 2021.
[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations (SimCLR)," in Proc. ICML, 2020.
[14] R. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers and Distillation through Attention (DeiT)," in Proc. ICML, 2021.
[15] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting," IEEE Trans. Biomed. Eng., vol. 67, no. 2, 2020.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. ICCV, 2017.
[17] X. Lu, Y. Zhu, S. Meng, and C. Zheng, "Self-supervised Vision Transformer for COVID-19 and Skin Cancer Screening," IEEE J. Biomed. Health Inform., vol. 26, no. 9, 2022.
[18] M. Hasan, S. Fatemi Shariatpanahi, and M. A. Al-Mamun, "DermNet: A Transfer Learning Approach for Skin Lesion Detection," Sensors, vol. 20, no. 18, 2020.
[19] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Melanoma Recognition in Dermoscopy Images via Aggregated Deep Convolutional Features," IEEE Trans. Biomed. Eng., vol. 66, no. 4, 2019.
[20] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marçal, and J. Rozeira, "PH2 — A Dermoscopic Image Database for Research and Benchmarking," in Proc. IEEE EMBC, pp. 5437–5440, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY