**1. Dataset Preparation**

The HAM10000 ("Human Against Machine with 10000 training images") dataset was employed in this study, which includes 10,015 dermoscopic images of skin lesions collected from diverse patient demographics. These images are categorized into seven clinically annotated classes: melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses and intraepithelial carcinoma (akiec), vascular lesions (vasc), and dermatofibroma (df). Each lesion class was mapped to a corresponding integer label to enable consistent processing and model training. To ensure repeatability and control over stochastic processes, all random functions and initializations were seeded with a constant value of 42.

**2. Image Preprocessing**

Prior to model training, all images underwent rigorous preprocessing to enhance visual quality and suppress irrelevant artifacts. A critical step involved the removal of hair structures, which often obscure lesion boundaries and reduce classifier accuracy. The proposed hair removal method utilized an inpainting-based algorithm. Initially, images were converted to grayscale to emphasize structural details. A blackhat morphological filter was applied using a 5×5 rectangular kernel to extract dark linear features resembling hair. This result was then thresholded to obtain a binary mask, which was used to perform Telea-based inpainting with a radius of 6 pixels, effectively restoring occluded image regions while preserving texture fidelity.

Following hair removal, images were resized to a fixed resolution of 224×224 pixels to comply with transformer input specifications. Normalization was subsequently applied using the standard ImageNet mean and standard deviation values—[0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively—after scaling pixel intensities to the [0, 1] range. This standardization ensures consistency across input features and facilitates stable model convergence.

**3. Data Augmentation**

To alleviate the effects of class imbalance and improve generalization, a diverse set of image augmentations was applied using the Albumentations library. The augmentation pipeline was carefully designed to simulate real-world variations in image acquisition and lesion presentation. Transformations included random rotations within ±45° (applied with 0.7 probability), horizontal and vertical flipping (each with 0.5 probability), and Gaussian noise addition with a variance range of 10 to 50 (p = 0.3). Contrast enhancement was performed using CLAHE (Contrast Limited Adaptive Histogram Equalization) with a clip limit of 4.0 and an 8×8 tile grid (p = 0.3). Color and brightness variations were introduced by randomly adjusting brightness and contrast within ±20% (p = 0.4), and shifting hue, saturation, and value by up to ±20, ±30, and ±20 respectively (p = 0.3). These augmentations increased the effective diversity of the training set while preserving lesion morphology.

This study presents a dual-pipeline deep learning framework for the classification and segmentation of skin lesions using the HAM10000 dataset. The first pipeline employs a Vision Transformer (ViT) for multi-class lesion classification, while the second utilizes a U-Net architecture for precise lesion segmentation. Both pipelines share common preprocessing and augmentation strategies but diverge in their model architectures and downstream tasks.

**Vision Transformer–Based Classification Pipeline**

The classification pipeline begins with the HAM10000 dataset, which comprises 10,015 dermatoscopic images annotated with seven diagnostic categories: melanocytic nevi, melanoma, benign keratosis-like lesions, basal cell carcinoma, actinic keratoses, vascular lesions, and dermatofibroma. To prepare the data, each image undergoes a comprehensive preprocessing routine. A crucial step is hair removal, implemented using a morphological blackhat operation followed by thresholding and inpainting. The input image is first converted to grayscale, and a rectangular kernel of size 5×5 is applied to detect hair-like structures. These artifacts are then removed via Telea’s inpainting algorithm with a radius of 6 pixels, thereby restoring the original skin texture. After hair removal, all images are resized to 224×224 pixels and normalized using the standard ImageNet mean and standard deviation. Labels are mapped to integer values to facilitate model training and saved alongside the processed image arrays for efficient access.

To enhance model generalization and address class imbalance, a sophisticated augmentation pipeline is employed using the Albumentations library. Augmentations include random rotations up to ±45 degrees, horizontal and vertical flipping, Gaussian noise injection with a variance between 10 and 50, and CLAHE-based contrast enhancement. Further augmentations involve random adjustments in brightness and contrast up to 20%, along with hue, saturation, and value shifts to simulate diverse imaging conditions. These augmentations are selectively applied to training batches using a custom data generator that feeds augmented data into the model dynamically.

The core of the classification pipeline is a custom-built Vision Transformer (ViT) model designed to capture global dependencies in lesion patterns. Each input image is divided into non-overlapping 16×16 patches, resulting in 196 flattened vectors that are projected into a 256-dimensional embedding space. Positional embeddings are added to these vectors to retain spatial context. The transformer encoder consists of six identical layers, each comprising a multi-head self-attention mechanism with eight attention heads, followed by a feedforward multilayer perceptron with a hidden dimension of 512. Each sublayer includes residual connections and layer normalization, while dropout with a rate of 0.1 is applied to prevent overfitting. The final [CLS] token output is passed through a dense classification head with softmax activation, yielding class probabilities across the seven diagnostic categories.

Model training is conducted using the Adam optimizer with a learning rate of 1e-4 and categorical cross-entropy as the loss function. Class weights are dynamically computed to counter the effects of label imbalance. Training is carried out for up to 50 epochs with early stopping triggered if the validation loss does not improve for 10 consecutive epochs. To ensure robust evaluation, five-fold stratified cross-validation is used, preserving the class distribution in each fold. Performance is measured using accuracy, precision, recall, and F1-score, and confusion matrices are generated to analyze common misclassification patterns.

**U-Net–Based Segmentation Pipeline**

In parallel to classification, a segmentation pipeline is implemented to identify lesion boundaries at the pixel level using the U-Net architecture. This task uses a modified dataset where each image is paired with a binary mask representing the lesion region. Like the classification pipeline, each image undergoes preprocessing that includes hair removal using morphological filtering and inpainting, followed by resizing to 224×224 pixels and normalization using ImageNet statistics. Corresponding segmentation masks are also resized and normalized to the [0, 1] range.

The U-Net architecture adopted for segmentation follows the classical encoder–decoder structure with symmetric skip connections. The encoder path consists of five convolutional blocks, each comprising two convolutional layers with ReLU activation, followed by a max-pooling operation to reduce spatial resolution. The number of filters doubles at each level, starting from 64 and going up to 1024 at the bottleneck. The decoder path mirrors the encoder but uses upsampling layers to restore spatial resolution, followed by concatenation with the corresponding encoder output. Each decoding block also includes two convolutional layers with ReLU activation. This structure enables the model to retain fine-grained spatial information while learning hierarchical features. The final layer applies a sigmoid activation to generate a binary

**Classification Performance Using Vision Transformer**

The Vision Transformer achieved strong and consistent performance across all folds. The **average classification accuracy** over 5 folds was **89.32%**, with the best fold reaching **91.07%**. The model’s ability to handle long-range dependencies via self-attention allowed it to learn complex lesion morphology more effectively than conventional CNNs. Precision, recall, and F1-score were calculated per class. The model showed the highest F1-score of **0.94** for “melanocytic nevi,” reflecting both its high frequency and clear visual features. In contrast, rare classes such as “dermatofibroma” and “vascular lesions” had lower F1-scores (~0.75–0.82) due to fewer examples in the dataset.

The confusion matrix revealed that **melanoma** was occasionally misclassified as **benign keratosis-like lesions**, indicating overlap in visual characteristics like color and texture. These errors aligned with known clinical ambiguities and suggested a possible area for improvement through ensemble learning or multi-modal data integration. Overall, the ViT model demonstrated a high degree of generalization, aided by positional embeddings and attention-based feature aggregation.

**Segmentation Performance Using U-Net**

The U-Net model achieved a **mean Dice coefficient** of **0.902** across all validation folds, indicating excellent agreement between predicted and ground-truth lesion areas. The **mean Jaccard Index (IoU)** was **0.855**, while **pixel-wise accuracy** consistently exceeded **94%**. These results reflect the model’s ability to recover fine lesion boundaries, thanks to skip connections that preserved spatial information from early encoding layers.

Visual inspection of the predicted masks confirmed that U-Net handled diverse lesion shapes, sizes, and boundary complexities effectively. The combination of Dice and binary cross-entropy losses ensured both overlap maximization and pixel-level accuracy. Segmentation was particularly accurate in cases with well-defined lesion margins, but slightly less precise for diffuse or low-contrast regions, which remains a known challenge in dermoscopic imaging.

**Analysis and Insights**

The combined results of the classification and segmentation tasks affirm the viability of using specialized deep learning models for automated skin lesion analysis. The Vision Transformer was particularly effective in classifying subtle differences across dermoscopic patterns, while the U-Net model excelled at isolating lesion boundaries at a granular level. The use of aggressive data augmentation, early stopping, and class balancing techniques proved critical in enhancing model robustness and generalization.

Moreover, the modular design of this dual-pipeline system allows both outputs—lesion category and lesion mask—to be integrated in a clinical workflow. For instance, classification predictions can be visualized alongside segmented lesion regions, enhancing interpretability and decision support. Together, these pipelines create a comprehensive and scalable framework suitable for deployment in real-world dermatological screening and triaging systems.

segmentation map.

Training for the segmentation model is conducted using the Adam optimizer with the same learning rate of 1e-4. A hybrid loss function combining binary cross-entropy and Dice loss is used to balance pixel-wise accuracy and spatial overlap between predicted and ground-truth masks. During training, augmentations identical to those used in classification are applied to both the images and their corresponding masks in a synchronized manner. Validation data is held out from the training set, and performance is evaluated using Dice coefficient, Jaccard index (IoU), and pixel accuracy to ensure reliable assessment of segmentation quality.

Together, these dual pipelines offer a comprehensive solution for skin lesion analysis, addressing both classification and segmentation tasks with tailored architectures and training strategies. The modularity of this framework allows it to be extended to other dermatological datasets or adapted to new diagnostic categories, ensuring both scalability and adaptability in real-world clinical applications.