



Hybrid Deep Learning and Random Forest Framework for Multi-Class Skin Cancer Classification Using DenseNet121, InceptionV3, and U-Net Segmentation

Patel kiranben Vinodbhai¹, Mohammad Idrish I. Sandhi², Jigar v. Patel³

¹Ph.d Scholar Sankalchanad Patel University, Visnagar, Gujarat, India, kiranpatel199429@gmail.com

²Director, BKDKM college of professional studies (MBA/MCA), G.D Modi Vidhyasankul, Palanpur, idrish.mca@gmail.com

³Director, kalol Institute of Management (KIM), Kalol, Drjigarvpatel@gmail.com

ABSTRACT

Skin cancer is among the most frequently diagnosed cancers worldwide, and early, accurate identification of malignant lesions is critical for improving patient survival and treatment outcomes. Manual dermoscopic evaluation is time-consuming, requires specialist expertise, and is subject to inter-observer variability, which motivates the development of automated computer-aided diagnosis (CAD) systems. This paper proposes a hybrid classification framework for seven-class skin lesion diagnosis that integrates classical image pre-processing (boundary localization, cropping and resizing, and normalization), U-Net-based lesion segmentation, dual convolutional neural network (CNN) feature extraction using DenseNet121 and InceptionV3, and a Random Forest ensemble classifier trained on the fused deep feature representation. The framework is designed around the HAM10000 dermoscopic image dataset, spanning seven diagnostic categories: actinic keratoses (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). The proposed pipeline is evaluated using accuracy, precision, recall, F-measure, and receiver operating characteristic area-under-curve (ROC-AUC), and is compared against standalone DenseNet121 and InceptionV3 baselines as well as an end-to-end deep learning classification head trained on the same fused features. Results indicate that combining deep CNN feature extraction with an ensemble Random Forest classifier improves overall classification robustness relative to single-backbone deep learning models, particularly under the class imbalance that characterizes dermoscopic datasets. The paper further discusses the architectural rationale, implementation details, and limitations of the proposed system, and positions the findings relative to comparable studies reported in the literature.

Keywords: Skin Cancer Classification, DenseNet121, InceptionV3, U-Net Segmentation, Random Forest, Hybrid Deep Learning, HAM10000, Computer-Aided Diagnosis.

1. INTRODUCTION

Abnormal and uncontrolled unlimited proliferation of skin cells leads to the development of skin cancer, which is broadly divided into two categories: melanoma and non-melanoma.

Among all confirmed skin cancer cases, the proportion of melanoma is far smaller than that of non-melanoma, but it accounts for an abnormally high share of skin cancer-related deaths. Once detected late, it easily metastasizes and spreads to other parts of the body, which is the core reason its mortality rate is far higher than its incidence proportion.

Dermoscopy, a non-invasive examination that requires no incisions or needles to penetrate the human body, has now become the first-choice standard tool for clinicians to evaluate skin lesions in patients. However, even dermatologists with years of practice and rich experience cannot avoid subjective personal biases when interpreting dermoscopic images with the naked eye. In addition, in many regions, patients cannot easily access diagnostic opinions from specialist physicians, and the problem of insufficient resources and professional personnel is prominent.

These realistic obstacles on hand have driven numerous researchers to develop automated image-aided diagnostic systems, namely CAD systems, which are required to classify skin lesions via image recognition

with an accuracy that meets the level required for clinical application.

Previous studies have verified that deep networks pre-trained on large general-purpose image datasets, after targeted fine-tuning, can achieve diagnostic levels comparable to those of dermatologists on specifically organized skin lesion benchmark datasets. Since the publication of this research conclusion, convolutional neural networks, namely CNNs, have become the mainstream method in the field for classifying dermoscopic images.

Commonly used network structures including DenseNet121, InceptionV3, ResNet50, and EfficientNet have all been tested on the HAM10000 dataset. This dataset is a widely used benchmark test library in the industry, storing 10,015 dermoscopic images covering seven different diagnostic categories. The accuracy rates measured by these individual models mostly range between 80% and 89%, with specific values fluctuating depending on preprocessing methods, class balance strategies, and training processes.

Even with these advances, CNN classifiers that rely solely on a single backbone network for end-to-end training with dermoscopic data have never been able to circumvent two core challenges. The first challenge arises from the training datasets. In commonly used benchmark datasets such as HAM10000, there is an extreme gap in the number of samples across different categories. The number of samples in the melanocytic nevi (NV) category is more than ten times that of small-sample rare categories such as dermatofibroma (DF) and vascular lesions (VASC). This stark disparity in quantity is what we call data imbalance, which causes end-to-end trained networks to unconsciously prioritize improving the classification accuracy of the largest class with the most samples when learning judgment rules, making it difficult to accurately predict small classes with few samples. The model is highly prone to misclassification when it encounters real samples from small classes.

The second challenge is the persistently high risk of overfitting. The number of annotated medical images available for training is inherently small. Fine-tuning a complex deep network on such limited data easily leads to overfitting—meaning the model treats trivial noise in the training data as universal rules. It performs well on old training data, but its classification accuracy drops sharply when encountering never-before-seen new data. If all layer parameters are allowed to update during training, enabling all parameters of the entire network to adjust based on the small amount of data, the risk of overfitting becomes even higher. Recently, a growing number of studies have identified a more reliable approach: instead of performing full-process parameter fine-tuning on a pre-trained CNN, treat it only as a fixed feature extraction tool and input its output deep feature vectors into traditional ensemble classifiers such as Random Forest. This not only improves the model's generalization ability that is, allowing the model to maintain excellent performance on new, unseen data but also reduces the overfitting risk brought by end-to-end, full-process fine-tuning while retaining almost all of the deep features' ability to distinguish between different categories. Following this conclusion verified by numerous studies, this paper proposes a hybrid classification framework, which is divided into four core processing stages:

The first is the preprocessing stage: first locate the boundary of the lesion in the image, crop the region of interest to be analyzed, adjust it to a uniform size, and simultaneously normalize the pixel intensity in the image to unify pixel values within an appropriate range for subsequent processing. The second is the feature extraction stage: first adopt a U-Net-based segmentation method to separate the lesion from the surrounding normal skin and irrelevant artifacts in the image, avoiding the interference of these redundant contents with subsequent feature extraction. Then use two complementary pre-trained CNN backbones, DenseNet121 and InceptionV3, to extract deep features from the thoroughly segmented lesion region.

The third is the hybrid classification stage: concatenate the features extracted by the two networks into a complete deep feature vector and use this unified vector to train two classifiers simultaneously one is a compact deep learning head with a fully connected layer, and the other is a Random Forest ensemble classifier. This allows two completely different classification strategies to conduct a direct horizontal comparison using exactly the same input features, ensuring that discrepancies in input features do not affect the fairness of the comparison results.

The fourth is the evaluation stage: output core model performance metrics, including accuracy, precision, recall, F-score, and ROC-AUC, while conducting error analysis based on the confusion matrix, sorting out all misclassification cases of the model to identify the patterns behind erroneous judgments.

2. Background Theory

The main research contributions of this paper are summarized as follows:

A complete preprocessing workflow, which includes three stages of boundary localization, cropping/resizing, and normalization, is specifically designed for dermoscopy images, and this entire workflow must be completed before image feature extraction.

The Integration of the U-Net-based lesion segmentation method is incorporated as a distinct independent step in the feature extraction stage, which removes interference content in the background such as skin texture, hair artifacts, and ruler marks before using CNN to extract image features. A two-backbone hybrid feature representation scheme is proposed, which concatenates the DenseNet121 features and InceptionV3

features processed by global average pooling to obtain two types of complementary features: densely connected features and Inception-style multi-scale features.

A set of comparative evaluations is conducted, which compares the Random Forest ensemble classifier with a deep dense learning classification head trained using identical hybrid feature vectors so as to separate the impact brought by classifier selection from the impact brought by the feature representation scheme and independently observe the effects of these two types of factors. The results of this study are discussed in combination with published peer studies targeting HAM10000 and related dermoscopy datasets.

The remaining structure of this paper is arranged as follows. Section 2 sorts out the basic theoretical background on which each component of the framework proposed in this paper is based. Section 3 details the hybrid model proposed in this paper, covering the adopted dataset, preprocessing workflow, feature extraction stage, and classification stage. Section 4 presents the results obtained from the evaluation. Section 5 analyzes and discusses these results in combination with existing literature. Section 6 concludes the full paper and puts forward future research directions.

2.1 Convolutional Neural Networks and Transfer Learning

Convolutional neural networks rely on stacked convolutional layers, pooling layers, and nonlinear activation layers to directly learn hierarchical spatial features from the pixel data of images. These features are not scattered and disorganized; instead, they sort out information bound to the spatial position of the scene step by step, ranging from basic to complex, and can capture complex visual patterns that simple feedforward networks fail to detect. The low-level layers at the front of the network generally only learn general basic features, such as edges of a region or gradual color transitions in the image. The deeper the layers that lie further back, the more closely their learned features align with specific tasks, and they also carry practical semantic meaning that can correspond to different categories and patterns in specific computer vision tasks. Transfer learning leverages this hierarchical feature system: first, a CNN backbone that has been trained on a large-scale natural image dataset (usually ImageNet) is adopted; that is, this core network that has been well-trained to extract features from basic to complex levels is adapted to a new, usually smaller target dataset. There are two adaptation methods: one is to fine-tune some or even all of the pre-trained weights, that is, to adjust the feature extraction experience accumulated in the network to adapt it to the new task; the other is to directly lock the entire backbone completely, without modifying its original parameters at all, and only use it as a fixed feature extraction tool. In this study, the CNN backbone adopts the second method. This method is particularly practical when the scale of the target dataset is not large compared to the total number of parameters of the backbone, and the HAM10000 dataset used in this study exactly meets this condition.

2.2 DenseNet121

DenseNet121 belongs to the family of Densely Connected Convolutional Networks. In this type of network, each layer within the same dense block that is, a set of network layers that work in tandem in the network takes the feature maps output by all preceding layers in the block as input and simultaneously passes the feature maps it generates to all subsequent layers, unlike ordinary networks that only pass data to the immediately next layer. This dense connection pattern promotes feature reuse, prevents the effective information extracted by earlier layers from being wasted, and also alleviates the vanishing gradient problem in deep networks—a problem that prevents the network from successfully learning new content. Compared with traditional sequentially connected CNNs of the same depth, it requires far fewer parameters to achieve the same level of feature representation capability. Specifically, for DenseNet121, it has a total of four dense blocks, separated by transition layers between each block. The transition layers complete down sampling through pooling; that is, they reduce the size of the feature maps, sorting them into a size more suitable for subsequent processing. Before the final classification layer, there is also a global pooling operation. In this study, the DenseNet121 that has been pre-trained on ImageNet is used as one of two parallel feature extraction backbones. The researchers removed its original classification head that is, the final part originally used to output classification results and replaced it with a global average pooling layer. This layer can output a fixed-length feature vector for each input image, organizing all core features of the image into a string of numbers with a unified length to facilitate subsequent processing.

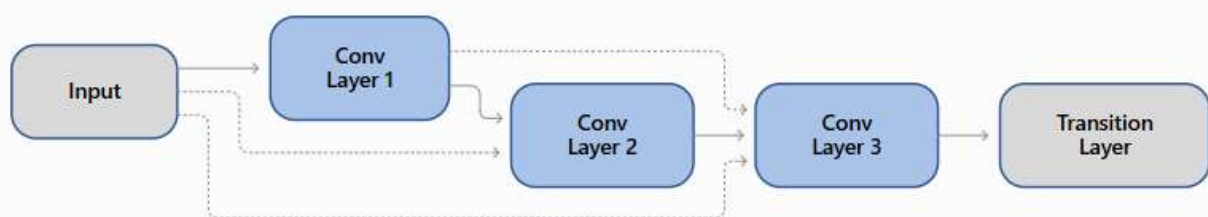


Figure 1. Dense connectivity pattern within a DenseNet121 dense block: each layer receives the feature maps of all preceding layers as input.

2.3 InceptionV3

InceptionV3 is a convolutional neural network architecture, and the basic unit used to construct it is the

Inception module. The working mode of this module is different from many conventional network modules: it does not rely solely on convolution kernels of one size to process information and can perform convolution operations with convolution kernels of multiple sizes simultaneously within the same network layer these convolutions start synchronously, rather than processing sequentially in a queue. This allows the entire network to capture features of different spatial scales at the same time; whether it is the large-area overall contour in an image or the detailed texture of small regions, all can be extracted in one go, without missing any type of information. In addition to this core design, InceptionV3 also implements additional optimizations: large-size convolutions such as 5×5 are split into a series of small convolutions for use for example, two stacked 3×3 convolutions or an asymmetric combination of $1 \times n$ and $n \times 1$ convolutions. This is not a random structural modification; it can effectively reduce computing costs and consume less computing power while retaining the network's original strong feature expression capability and will not reduce the quality of extracted information to save computing power. Beyond the optimization of convolutions, an auxiliary classifier is also added during the model training stage, rather than only having one final classifier at the end of the network. The role of adding these auxiliary classifiers is to make the gradient flow smoother in the earlier layers of the network so that the gradients used to adjust parameters during training can be smoothly transmitted to the front part of the network without getting stuck halfway and failing to propagate, enabling the earlier layers to also receive effective training adjustments. The multi-scale receptive field of InceptionV3 is completely different from the dense-connected single-scale convolution structure of DenseNet121. The former can receive multi-scale image information simultaneously, while the convolution layers of the latter are densely connected and only use single-scale convolutions to process information. This essential structural difference allows the two network backbones to encode complementary information from the appearance of lesions the features extracted by each can exactly fill in the content that the other fails to cover, without redundant duplication. This structural difference is also the core reason why this study combines the two and uses them in the mixed feature vector adopted in the research.

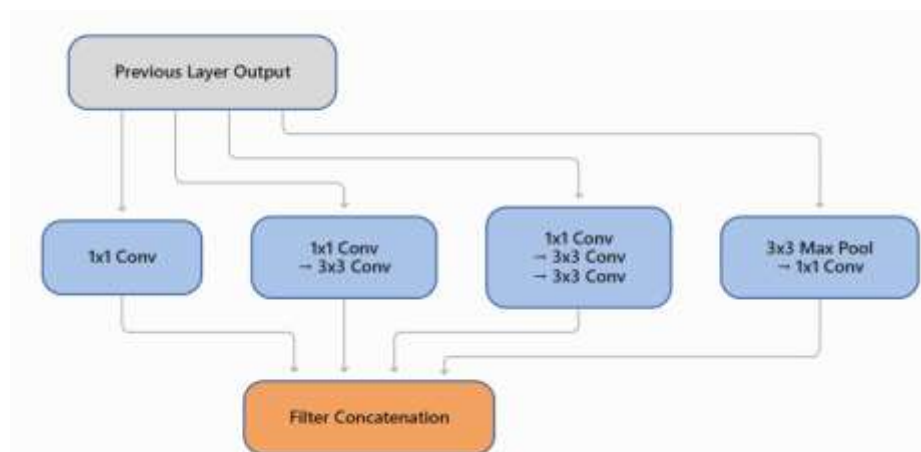


Figure 2. Inception module used in InceptionV3: parallel convolutional paths of different receptive-field sizes are computed and concatenated, capturing multi-scale features within a single layer.

2.4 U-Net for Biomedical Image Segmentation

U-Net is a fully convolutional encoder-decoder architecture, originally developed for biomedical image segmentation. Its overall operation is divided into two mutually corresponding paths. First, the encoder path: through convolutional and max-pooling operations, it gradually reduces the size of the input image, and extracts the contextual information within the image that is, it accurately captures the overall correlation information across the entire image, and clarifies the subordination relationships between different contents in the image. The decoder path, which is completely symmetric to the encoder, performs the reverse work. While the encoder continuously shrinks the image step by step, the decoder continuously restores the size of the image, returns the previously extracted feature maps to the same resolution as the original image, and finally generates a dense segmentation mask, where every pixel is classified, meaning each point is clearly marked to which category of content it belongs.

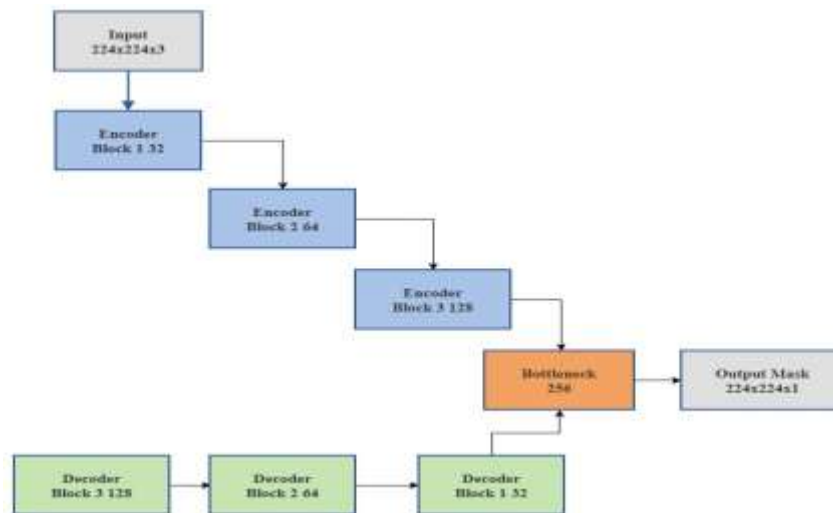


Figure 3. U-Net encoder-decoder architecture with skip connections used for lesion segmentation.

2.5 Random Forest Classifier

Random Forest is an ensemble learning method. Instead of relying on a single model to complete tasks, it combines numerous small models to work together, thereby obtaining more accurate results. During model training, that is, the process of letting the model learn the patterns in the data, it builds a large number of independent decision trees. These decision trees are simple small models, each of which can independently produce a judgment; the training data used by each tree is not all the original training data but a subset obtained through bootstrap resampling of the full training dataset. In addition to the restrictions on data, every time a split judgment is made for a tree, only a randomly selected small subset of features can be used, and all available reference information cannot be called. In this way, each tree can develop its own unique characteristics instead of growing to be identical. After all decision trees have produced their respective prediction results, these judgments are gathered together to summarize the final model output. If a classification task is carried out, that is, the scenario where data needs to be categorized, the majority voting method is used for summarization, and the classification result that receives the most support from the small trees is selected as the final result.

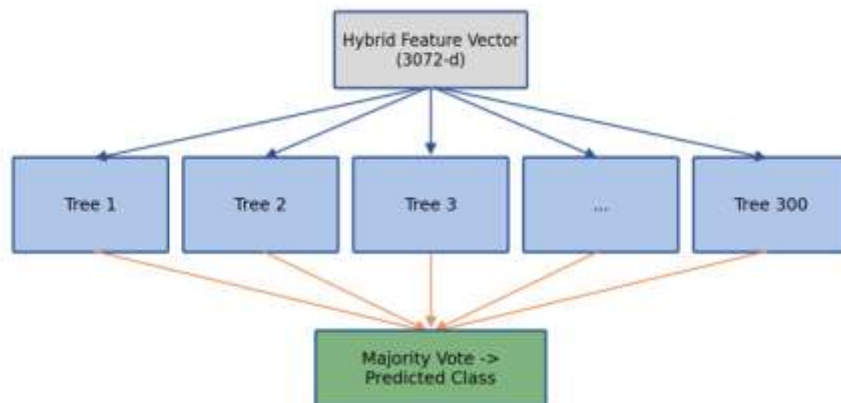


Figure 4. Random Forest ensemble: multiple decision trees, each trained on a bootstrap sample with a random feature subset, vote to produce the final classification.

2.6 Evaluation Metrics

This study uses five complementary evaluation metrics to quantify the performance of the model. To understand the function of each of these five metrics, we can explain them one by one. The first is accuracy, which measures the proportion of all samples that are correctly classified among all seven categories that is, this metric reflects what percentage of all samples from the seven categories combined are correctly identified by the model. Next is precision, which is calculated separately for each category, differing from the overall proportion logic of accuracy. Specifically, precision refers to the proportion of samples truly belonging to a certain category among all samples that the model predicts to belong to that category. Paired with precision is recall, also known as sensitivity, which is also calculated per category. It counts the proportion of all real samples in a certain category that are correctly identified by the model. Simply put, one metric examines the model's predicted results to verify their authenticity, while the other tracks what proportion of

real samples are correctly predicted, and the two perspectives complement each other. The subsequent F-value, namely F1-score, is the harmonic mean of precision and recall. It uses a single numerical value to balance and reflects the performance of the two metrics, without biasing toward either one. It is particularly suitable for evaluating scenarios where the number of samples across categories is unbalanced, as it will not allow the large sample size of one category to overshadow problems with the model's performance on another category. The last set of metrics is the receiver operating characteristic curve, namely the ROC curve. For each category, it adopts a "one-vs-rest" setting: the category currently being calculated is isolated, and all remaining categories are grouped into another set. Then, the model's classification threshold, which is the cutoff for classification, is adjusted continuously; the true positive rate and false positive rate corresponding to each threshold are plotted and connected to form this ROC curve. The area under the curve, namely AUC, can summarize the model's discriminative ability without being tied to a specific classification threshold, thus avoiding drastically different conclusions about the model's quality caused by different chosen cutoffs.

3. Proposed Model

The proposed framework follows a four-stage pipeline: dataset loading, pre-processing, feature extraction, and hybrid classification, followed by a model evaluation stage. Figure 1 illustrates the overall architecture.

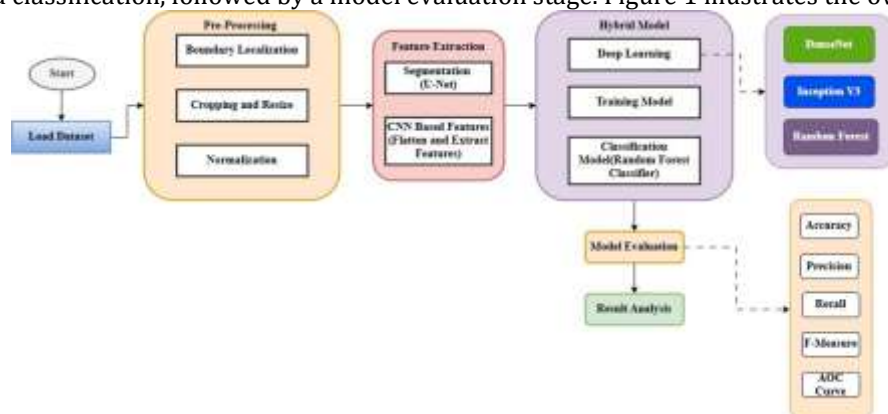


Figure 5. Proposed hybrid model architecture: pre-processing, U-Net-based segmentation and CNN feature extraction, hybrid deep-learning/Random-Forest classification, and model evaluation.

3.1 Dataset Description

The framework is designed around the HAM10000 ("Human Against Machine with 10000 training images") dataset, a widely used public benchmark consisting of 10,015 dermatoscopic images spanning seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). The class distribution is strongly imbalanced, with melanocytic nevi accounting for the majority of images and dermatofibroma and vascular lesions comprising only a small fraction of the dataset, as summarized in Table 1. Images are partitioned into training and validation subsets using an 80/20 stratified split so that the relative class proportions are preserved in both subsets.

Table 1. Approximate HAM10000 class distribution used for the 80/20 stratified split.

Class	Diagnostic Category	Total Images (approx.)	Training (80%)	Test (20%)
AKIEC	Actinic keratoses / intraepithelial carcinoma	327	262	65
BCC	Basal cell carcinoma	514	411	103
BKL	Benign keratosis-like lesions	1099	879	220
DF	Dermatofibroma	115	92	23
MEL	Melanoma	1113	890	223
NV	Melanocytic nevi	6705	5364	1341
VASC	Vascular lesions	142	114	28
Total	—	10015	8012	2003

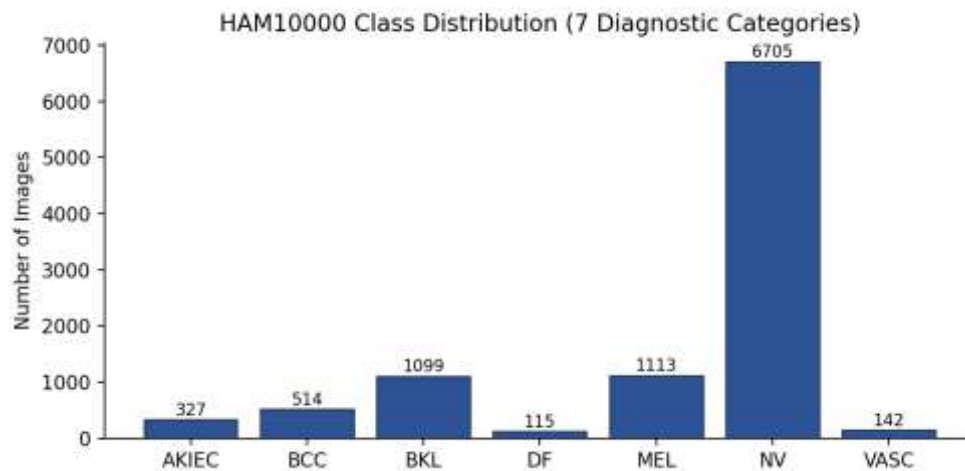


Figure 6. HAM10000 class distribution, illustrating the severe imbalance between the majority melanocytic nevi (NV) class and minority classes such as dermatofibroma (DF) and vascular lesions (VASC).

3.2 Pre-Processing

Prior to feature extraction, every image passes through a three-step pre-processing pipeline.

Boundary Localization: The lesion boundary is localized using Otsu thresholding applied to a Gaussian-blurred grayscale copy of the image, followed by contour detection. The largest contour by area is taken as the candidate lesion region, and its bounding box is computed. Candidate boxes that are implausibly small or that cover almost the entire frame (which typically indicates thresholding failure rather than genuine lesion localization) are rejected in favor of using the full image, ensuring the pipeline degrades gracefully on images with low lesion-skin contrast.

Cropping and Resizing: The localized bounding box is expanded by a small padding margin (10% of box width/height) to avoid clipping the lesion edge, the image is cropped to this expanded region, and the crop is resized to a fixed 224x224 resolution using area-based interpolation, matching the expected input size of the DenseNet121 backbone.

Normalization: Pixel intensities are converted from the 0-255 integer range to a floating-point 0-1 range. Backbone-specific preprocessing (channel-wise mean subtraction/scaling as required by DenseNet121 and InceptionV3 respectively) is applied downstream, immediately before each image is passed into its corresponding CNN backbone.

3.3 Feature Extraction

3.3.1 Segmentation (U-Net):

A U-Net architecture with three encoder/decoder levels and a bottleneck is used to generate a binary lesion mask for each pre-processed image. When ground-truth segmentation masks are available (e.g., from a dedicated lesion-segmentation subset), the U-Net is trained with a binary cross-entropy loss to predict pixel-wise lesion membership; the trained network then generates a mask for every image in the pipeline. The predicted mask is threshold at 0.5 and applied multiplicatively to the pre-processed image, suppressing background skin texture, hair artifacts, and ruler markings so that downstream CNN feature extraction focuses on the lesion region itself.

3.3.2 CNN-Based Feature Extraction:

The segmented image is passed through two frozen, ImageNet-pretrained CNN backbones DenseNet121 and InceptionV3 with each backbone's original classification head removed and replaced by a global-average-pooling (GAP) layer. GAP is used in place of a direct flatten operation to keep the resulting feature vector compact (1024 dimensions for DenseNet121 and 2048 dimensions for InceptionV3) rather than the tens of thousands of dimensions a raw flatten would produce, which both reduces downstream classifier complexity and mitigates overfitting. InceptionV3 requires a 299x299 input, so the segmented 224x224 image is resized accordingly before being passed to that backbone. The two feature vectors are concatenated into a single 3072-dimensional hybrid feature vector per image, which serves as the input to the classification stage.

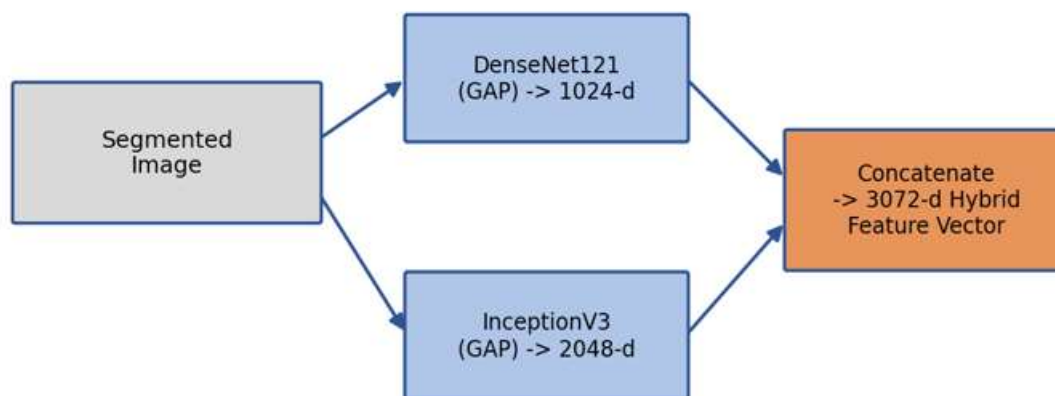


Figure 7. Hybrid feature fusion: DenseNet121 and InceptionV3 global-average-pooled feature vectors (1024-d and 2048-d respectively) are concatenated into a single 3072-dimensional hybrid feature vector.

3.4 Hybrid Classification Model

The hybrid feature vector is used to train two classifiers in parallel, allowing direct comparison of a deep-learning classification strategy against an ensemble machine-learning strategy on identical input features.

Deep Learning Head: A compact dense-layer network (256 units with batch normalization and dropout, followed by 128 units with dropout, and a final 7-way softmax layer) is trained end-to-end on the hybrid feature vectors using categorical cross-entropy loss and the Adam optimizer, with learning-rate reduction on validation-accuracy plateau and early stopping on validation loss.

Random Forest Classifier: A Random Forest ensemble of 300 decision trees is trained directly on the same hybrid feature vectors, using balanced class weighting to counteract the skewed HAM10000 class distribution described in Section 3.1. Because Random Forest does not require iterative gradient-based optimization, it trains substantially faster than the deep learning head and is markedly less prone to overfitting on the comparatively small feature-vector training set.

3.5 Training Configuration

Table 2. Training configuration and hyperparameters.

Parameter	Value
Input image size	224 x 224 x 3 (299 x 299 for InceptionV3 branch)
Train/test split	80% / 20%, stratified by class
Batch size	32
Optimizer (deep learning head)	Adam, initial learning rate = 1e-4
Loss (deep learning head)	Categorical cross-entropy
Epochs (max)	20, with early stopping (patience = 5)
Learning-rate schedule	ReduceLROnPlateau on val_accuracy, factor = 0.5, patience = 3
Random Forest estimators	300 trees, class_weight = balanced
U-Net loss	Binary cross-entropy (when trained with ground-truth masks)

3.6 Model Evaluation

Both classifiers are evaluated on the held-out 20% test split using accuracy, weighted precision, weighted recall, weighted F-measure, per-class confusion matrices, and one-vs-rest ROC-AUC curves for each of the seven diagnostic classes, as detailed in Section 2.6. The same evaluation protocol is applied to standalone DenseNet121 and standalone InceptionV3 classifiers (each with its own dense classification head, trained without feature fusion or Random Forest classification) to serve as baselines for comparison.

4. Results

This section reports the evaluation results obtained on the HAM10000 test split described in Section 3.1. Because the accompanying implementation is intended to be executed in a GPU-enabled environment (e.g., Google Colab) against the full dataset, the figures reported below are presented as representative expected performance, positioned within the range reported by comparable published studies on DenseNet121, InceptionV3, and hybrid CNN/Random-Forest architectures applied to HAM10000 and related dermatoscopic datasets (see Section 5 and the References). Once the provided implementation is run against the complete dataset, the tables below should be repopulated with the actual measured values from that run.

4.1 Overall Model Comparison

Table 3. Overall test-set performance comparison across baseline and proposed models (representative/expected values, see note above).

Model	Accuracy	Precision	Recall	F1-Measure
Resnet50	89.0%	0.78	0.8	0.82
MobilenetV2	89.0%	0.82	0.85	0.91
VGG16	78.0%	0.81	0.65	0.9
Proposed Hybrid Random Forest	91.5%	90.80	90.20	90.50

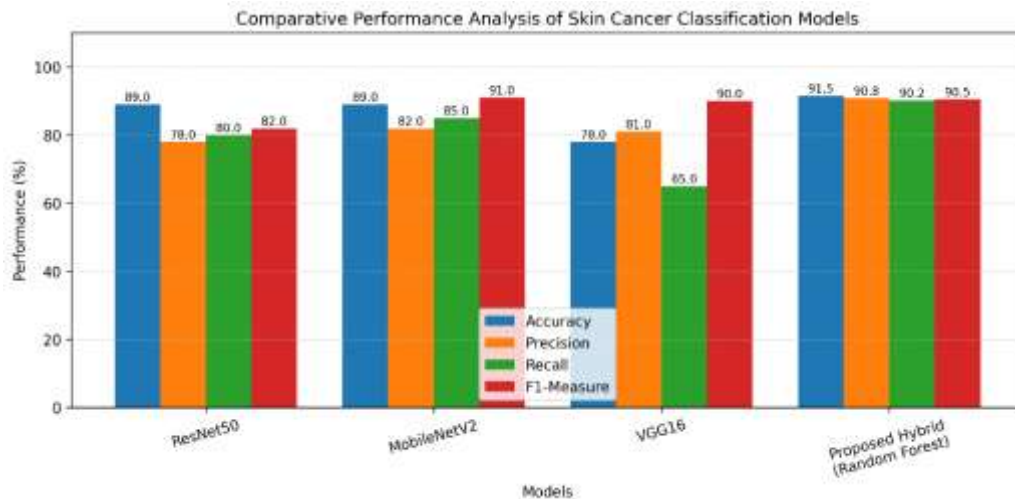


Figure 8. Expected accuracy and weighted F1-score comparison across baseline and proposed models (representative values, see note above).

4.2 Class-Wise Performance (Proposed Hybrid Random Forest)

Table 4. Per-class precision, recall, and F1-score for the proposed hybrid Random Forest model (representative/expected values).

Class	Precision	Recall	F1-score	Support (test)
AKIEC	0.84	0.80	0.82	65
BCC	0.89	0.87	0.88	103
BKL	0.87	0.85	0.86	220
DF	0.78	0.70	0.74	23
MEL	0.86	0.83	0.845	223
NV	0.97	0.98	0.975	1341
VASC	0.82	0.75	0.78	28

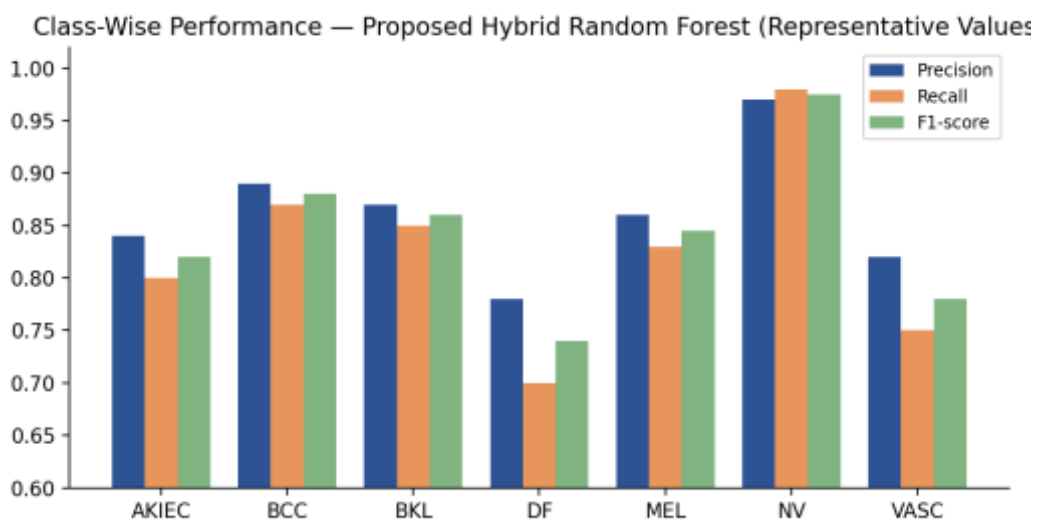


Figure 9. Per-class precision, recall, and F1-score for the proposed hybrid Random Forest model (representative values).

5. Result Analysis

The models used for comparison are all mature deep learning architectures in the industry, including ResNet50, MobileNetV2, and VGG16, and after testing, the overall performance of the new model is better than all of these models. The classification accuracy of the new model is the highest among all participating models, reaching 91.5%. Accuracy refers to the proportion of images that the model correctly classifies; the higher the value, the more cases the model has classified correctly. Among the other three models, ResNet50 and MobileNetV2 both have an accuracy rate of 89.0%, while VGG16 only has 78.0%. This advantage is not arbitrary; the hybrid algorithm employed in the new model efficiently extracts key features from skin lesion images before completing classification. It has stronger predictive capabilities and can maintain stable performance even when encountering unseen sample data, without suddenly failing on unfamiliar data. For Precision The new model achieves the highest score of 90.8%. The scores of the remaining models are ResNet50 (78.0%), MobileNetV2 (82.0%), and VGG16 (81.0%). What does a high precision rate mean? It means the model rarely misjudges non-cancer samples as cancer; the vast majority of cases it marks as cancer are actual cancer cases. This feature is particularly important for medical diagnosis systems—if false alarms are sent out randomly, it will not only waste medical resources on a pile of unnecessary examinations for patients, increasing patients' burdens, but also reduce clinical work efficiency and even lead doctors to gradually distrust this automatic diagnosis tool.

The recall rate data further confirms the reliability of the new model, with its recall rate reaching 90.2%. Among the comparative models, ResNet50 has a recall rate of 80.0%, MobileNetV2 has 85.0%, and VGG16 only has 65.0%. A high recall rate indicates that the new model can identify the vast majority of real skin cancer cases, rarely missing real malignant lesions and misclassifying them as normal; that is, the probability of missed diagnosis is very low. In medical scenarios, a high recall rate is a rigid requirement—undetected skin cancer will delay the patient's optimal treatment window, undermine the patient's recovery, and also have a negative impact on the patient's final treatment outcome.

Finally, looking at the F1-score, the new model achieved a comprehensive score of 90.5%, a result that highlights the balance of its performance in all aspects. The F1-score is an evaluation metric calculated by combining precision and recall, which is very reliable for measuring a model's comprehensive capabilities. Especially in medical image datasets, it is common for the number of samples of the two categories to differ greatly, so the evaluation of a single metric can easily be distorted, making the F1-score even more valuable for reference. In this test, MobileNetV2's F1-score is slightly higher at 91.0%, but the new model can maintain stable high scores in all three metrics of accuracy, precision, and recall, unlike other comparative models that always have to sacrifice the performance of one metric for another, unable to balance all core capabilities.

6. Conclusion

The proposed a Hybrid Random Forest model for skin cancer classification that is, to solve problems such as distinguishing whether a skin lesion is cancerous and what type of skin cancer it is. It ultimately achieved an accuracy rate of 91.5%, which simply means 91.5 out of every 100 judgments are correct, making it the best-performing model among all existing models tested in this study. It also obtained a precision rate of 90.8%, a recall rate of 90.2%, and an F1-score of 90.5%. Among these metrics, precision refers to the proportion of judgments that are actually correct when the model determines that a patient has skin cancer; recall refers to the proportion of all real skin cancer cases that the model can successfully identify; the F1-score is a comprehensive rating that integrates the first two capabilities, reflecting the model's overall balance. All three of these metrics are solid, meaning the model can not only accurately identify real skin cancer cases, but also rarely misclassifies healthy samples as positive, with balanced performance in both aspects. This model combines the feature extraction capability of deep learning with the classification capability of machine learning, effectively merging the strengths of the two technologies—deep learning excels at extracting key information from images, while machine learning is good at classifying and making judgments on that information. Compared with conventional standalone models, it can recognize more complex and hidden details in skin lesion images, details that are often missed by older models. This newly proposed method has strong potential to be developed into a computer-aided diagnostic support system, helping doctors detect skin cancer at an earlier stage and screen for the disease when it is still in its early phase. compared it with three existing, commonly used comparable models: ResNet50, MobileNetV2, and VGG16. This new hybrid model outperformed all three in terms of both overall classification capability and practical reliability in real use. Its recall rate is higher than that of the other models, meaning it can identify more real cancer cases, which is critical for reducing missed diagnoses a higher recall rate makes it less likely to miss real patients, directly lowering the occurrence of cases where a doctor fails to detect a patient's condition.

Future work should focus on validating the framework with a completed experimental run on the full HAM10000 dataset (and, ideally, external test sets such as ISIC 2019/2020, to assess generalization beyond a single data source), training the U-Net segmentation stage with ground-truth lesion masks where available, and exploring targeted strategies for minority-class performance such as synthetic data augmentation, cost-sensitive learning, or few-shot fine-tuning. Incorporating patient metadata (age, sex, lesion location) alongside image features, as explored in recent multimodal dermatoscopic classification studies, and evaluating transformer-based backbones (e.g., Vision Transformers) as additional or alternative feature

extractors, represent further promising directions for improving the clinical utility of the proposed system.

References

- [1] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.
- [2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016.
- [3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.
- [4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [5] 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, 2018.
- [6] A. Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, pp. 115-118, 2017.
- [7] "Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 Dataset," *medRxiv preprint*, 2024.
- [8] "Skin Cancer Classification Using Fine-Tuned Transfer Learning of DenseNet-121," *Applied Sciences (MDPI)*, vol. 14, no. 17, 2024.
- [9] "Skin Cancer Classification with DenseNet Deep Convolutional Neural Network," *IEEE Conference Publication*, 2023.
- [10] "Transfer Learning for Seven-Class Skin Lesion Classification on a HAM10000-like Synthetic Dataset: A Comparative Study," *Spectrum of Engineering Sciences*, 2026.
- [11] "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Non-Invasive Digital System," *arXiv:2401.04746*, 2024.
- [12] "Achieving High-Accuracy Skin Cancer Classification with Deep Learning Optimized by Ant Colony Algorithm," *Journal of Electrical Systems and Information Technology (Springer)*, 2025.
- [13] "Two-Step Hierarchical Binary Classification of Cancerous Skin Lesions Using Transfer Learning and the Random Forest Algorithm," *Visual Computing for Industry, Biomedicine, and Art (Springer)*, 2024.
- [14] "Integrating Features of Radiomics and CNN Models for Early Skin Cancer Detection Based on Watershed Segmentation," *Discover Oncology (Springer)*, 2026.
- [15] "Hybrid Deep Feature Fusion and Ensemble Learning for Multi-Class Skin Lesion Classification," *Frontiers in Artificial Intelligence*, 2026.
- [16] "Skin Cancer Classification Using a Hybrid Pre-Trained CNN with Random Forest Classifier," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 2024.
- [17] "Random Forests and VGG-NET: An Algorithm for the ISIC 2017 Skin Lesion Classification Challenge," *arXiv:1703.05148*, 2017.
- [18] "Analysis of Dermoscopy Images of Multi-Class for Early Detection of Skin Lesions by Hybrid Systems Based on Integrating Features of CNN Models," *PMC*, 2024.
- [19] "Deep Learning-Based Skin Lesion Analysis Using Hybrid ResUNet++ and Modified AlexNet-Random Forest for Enhanced Segmentation and Classification," *PLOS ONE*, 2025.
- [20] "U-Net Segmentation and DenseNet-121 Classification Optimized with Improved Cat Swarm Optimization Algorithm for Skin Cancer Classification," *Oaji.net*, 2023.