Towards Accurate and Explainable Skin Cancer Identification: A Comparative Analysis of Machine Learning, Deep Learning, Vision Transformers and Hybrid XAI-Based Methods

Mr. Anshid Babu T M[1], Assistant Professor, Department of Computer Science and Artificial Intelligence, SAFI Institute of Advanced Study Vazhayur (Autonomous), Malappuram.

Dr. Arshad P T [2], Assistant Professor , Department of Computer Science and Artificial Intelligence ,

SAFI Institute of Advanced Study(Autonomous) Vazhayur , Malappuram

Mr. Abil Antony[3], Assistant Professor, Department of MCA, KMM College of Arts and Science, Thrikkakara, Ernakulam

Ms. Vishnupriya Mohan[4], Assistant Professor, PG department of Computer Science,

St. Peter's Kolenchery, Ernakulam

Abstract

Skin cancer identification from dermoscopic images has become an important application of computer vision and artificial intelligence. The visual similarity between several benign and malignant lesions makes automated classification a challenging task, particularly when images contain hair, illumination variations, acquisition artifacts, and substantial differences in lesion appearance. Machine-learning and deep-learning techniques have consequently been investigated as computer-assisted tools for supporting dermatological image analysis. However, reported performance varies significantly between studies because researchers frequently employ different datasets, preprocessing techniques, class-balancing strategies, network architectures, data partitions, and evaluation metrics. Therefore, a direct and fair comparison of existing approaches is required.

This paper presents a comparative study of conventional machine-learning algorithms, convolutional neural networks, transfer-learning models, ensemble methods, and transformer-based architectures for multi-class skin lesion identification. The study considers the HAM10000 dataset as the primary benchmark and proposes ISIC data for external evaluation. A new framework named the Explainable Hybrid Skin Cancer Identification Framework (EHSC-XAI) is introduced. The framework combines lesion-focused preprocessing, artifact reduction, class-aware augmentation, multi-scale deep feature extraction, attention-based feature fusion, metadata integration, ensemble prediction, confidence calibration, and explainable artificial intelligence. The proposed design combines complementary visual representations rather than depending on a single classifier.

An important feature of the framework is its emphasis on reliability rather than accuracy alone. Grad-CAM, Grad-CAM++, LIME, and SHAP are incorporated to investigate the regions and features influencing individual predictions. In addition, a confidence-aware decision layer is proposed so that highly uncertain images can be referred for further clinical assessment instead of being assigned an apparently certain diagnosis. The paper also proposes an ablation study to determine the contribution of preprocessing, augmentation, attention, metadata, and ensemble learning independently.

The comparative analysis indicates that deep transfer-learning and ensemble approaches generally provide stronger representations than traditional handcrafted-feature classifiers, while transformer architectures provide an alternative mechanism for capturing long-range visual relationships. Nevertheless, performance reported across independent studies cannot be interpreted as a universal ranking because of differences in experimental conditions. The proposed framework therefore emphasizes standardized evaluation using macro-F1, balanced accuracy, sensitivity, specificity, ROC-AUC,

PR-AUC, calibration error, computational cost, and explanation quality. The study provides a research direction toward skin lesion classification systems that are not only accurate but also interpretable, uncertainty-aware, and suitable for future external clinical validation.

Keywords: Machine Learning, Skin Cancer, Melanoma, Skin Lesion Classification, Dermoscopy, HAM10000, ISIC, Convolutional Neural Network (CNN), Transfer Learning, Ensemble Learning, Vision Transformer (ViT), Explainable Artificial Intelligence (XAI), Grad-CAM, LIME, SHAP.

1. Introduction

Skin cancer is an important area of medical diagnosis in which early recognition can influence subsequent clinical management. A skin lesion may have a relatively harmless appearance while another visually similar lesion may require immediate clinical attention. This similarity creates difficulty even when images are examined using dermoscopy, particularly for lesions with unusual colour distribution, irregular boundaries, heterogeneous texture, or limited visual contrast.

Traditional diagnosis is primarily dependent on clinical examination, dermoscopic assessment, patient information, and, when required, histopathological confirmation. Computer-assisted analysis is not intended to replace these procedures. Instead, artificial intelligence can potentially provide an additional source of information that assists specialists in screening, prioritization, and second-opinion workflows.

The availability of public dermoscopic image repositories has accelerated research in this area. One of the most widely used resources is HAM10000, which contains 10,015 dermoscopic images collected from multiple sources and covering several common pigmented lesion categories. The International Skin Imaging Collaboration has also organized several public challenges for lesion segmentation, feature analysis, and classification. The ISIC 2019 challenge provides 25,331 training images and addresses multiple diagnostic categories, with an additional outlier category considered in the challenge testing framework.

The early generation of automated skin lesion systems generally depended on manually designed descriptors. Colour, texture, shape, border irregularity, symmetry, and other image characteristics were extracted and supplied to conventional classifiers. Such systems remain useful for establishing baseline performance, but manually selected descriptors may not capture the full visual complexity of dermoscopic lesions.

Deep learning changed this situation by allowing models to learn image representations directly from training data. Convolutional neural networks can identify increasingly complex patterns through successive layers. Lower layers may learn edges and colour transitions, intermediate layers can represent texture and local structures, and deeper layers can capture more abstract visual patterns.

Transfer learning has become especially useful because medical datasets are generally smaller than the datasets used to train large general-purpose vision models. A pretrained network can be adapted to the skin-lesion domain by replacing its final classification layer and fine-tuning selected layers.

However, a high classification accuracy does not automatically indicate that a system is clinically reliable. A dataset can contain a large majority class, allowing a model to obtain a respectable overall accuracy while performing poorly on less frequent but clinically important classes. Dataset-specific characteristics can also produce artificially optimistic results when training and testing images are not sufficiently independent.

Another major issue is interpretability. A conventional classifier can produce a prediction such as “melanoma” without explaining why the prediction was generated. This is problematic when an AI system is intended to assist medical professionals. Explainable artificial intelligence can provide visual or feature-based evidence about the model's decision. Current research frequently examines methods such as Grad-CAM, LIME, and SHAP, but recent comparative work has

also highlighted that an explanation should not automatically be considered clinically meaningful simply because a heatmap is available.

Therefore, this study focuses on four interconnected questions:

1. How do conventional machine-learning algorithms compare with modern deep-learning approaches?
2. Which preprocessing and learning strategies are most useful for improving classification?
3. Can combining complementary CNN and transformer representations provide a more robust classifier?
4. Can explainable AI and confidence estimation make predictions more transparent and safer?

To address these questions, this paper proposes the **Explainable Hybrid Skin Cancer Identification Framework (EHSC-XAI)**.

Main contributions

The main contributions of the study are:

- A unified comparison of traditional ML, CNN, transfer-learning, ensemble, and transformer approaches.
- A structured preprocessing pipeline for dermoscopic images.
- A class-aware augmentation strategy designed for imbalanced lesion categories.
- A hybrid feature representation using complementary deep-learning architectures.
- Integration of available clinical metadata with image-derived features.
- An ensemble classification mechanism.
- A confidence-aware prediction layer for uncertain cases.
- Multi-method XAI using Grad-CAM, Grad-CAM++, LIME, and SHAP.
- A proposed explanation-consistency measure for comparing different XAI outputs.
- An ablation framework for determining the actual contribution of each proposed component.
- A methodology emphasizing external validation rather than relying exclusively on a single benchmark dataset.

2. Literature Review

The development of automated skin lesion classification has progressed through several stages, beginning with handcrafted image descriptors and conventional machine-learning algorithms and gradually moving toward deep neural networks, transfer-learning architectures, ensemble strategies, and transformer-based models. Each stage has addressed specific limitations of the preceding approaches. However, improvements in computational performance have also introduced new challenges, including increased model complexity, data requirements, class imbalance, computational cost, and reduced transparency of prediction mechanisms. Understanding this progression is important for establishing an appropriate comparative framework for automated skin cancer identification.

The earliest computer-assisted approaches primarily represented a skin lesion through manually selected visual characteristics. Since a dermoscopic image contains information related to colour, texture, shape, and boundary structure, researchers developed numerical descriptors to represent these properties. Colour-based representations can be obtained from spaces such as RGB, HSV, or other transformed colour domains. Texture information can be characterized using statistical measurements, local neighbourhood patterns, or frequency-domain representations. Similarly, geometric descriptors can capture properties such as lesion area, perimeter, circularity, asymmetry, compactness, and boundary irregularity. These extracted characteristics are then supplied to a classification algorithm to distinguish between different lesion categories. Such approaches have the advantage of being relatively simple to implement and can operate with considerably fewer computational resources than modern deep-learning systems. However, their performance is strongly influenced by the quality of the manually selected features. A feature that

captures a useful characteristic under one imaging condition may become less effective when illumination, acquisition equipment, skin appearance, or lesion characteristics change.

Among conventional classifiers, the Support Vector Machine (SVM) has been widely used for high-dimensional classification because it attempts to identify an effective decision boundary between classes and can employ nonlinear kernel functions when the original feature space is insufficient for separation [1], [2]. Random Forest (RF) provides another useful approach by combining decisions from multiple decision trees, allowing nonlinear relationships between extracted features and diagnostic categories to be represented [3]. K-Nearest Neighbour (KNN) performs classification according to the similarity between a new observation and previously labelled samples, while logistic regression provides a relatively simple probabilistic baseline for classification [4]. Although these methods remain valuable as benchmark models, their dependence on manually constructed representations limits their ability to discover complex visual patterns that may not be obvious during feature-design.

The emergence of Convolutional Neural Networks (CNNs) introduced a major change in image-based skin lesion analysis. Rather than requiring a researcher to define colour, texture, or shape descriptors explicitly, a CNN can learn a hierarchy of representations directly from image data. During the initial stages of processing, convolutional layers may capture simple visual structures such as edges, colour transitions, and local patterns. Deeper layers progressively combine these representations to describe more complex structures associated with lesions. This hierarchical learning mechanism makes CNNs particularly suitable for image classification and has contributed to their widespread adoption in medical-image analysis [5], [6].

Several CNN architectures have been investigated for skin lesion classification, with each architecture offering a different strategy for learning visual representations. Residual networks introduce shortcut connections that allow information to move across multiple layers and help in optimizing deep architectures [7]. Dense convolutional networks establish extensive connections between successive layers, encouraging the reuse of previously learned information [8]. Inception-based networks process information using multiple receptive-field scales, allowing visual patterns of different sizes to be considered within the same architecture [9]. Xception modifies the conventional convolutional operation through depthwise separable convolutions, reducing the computational burden while maintaining a strong capacity for visual representation [10]. EfficientNet further investigates systematic scaling of network dimensions, providing a family of models that balance predictive capability with computational requirements [11].

The increasing use of CNNs in skin lesion analysis has also exposed an important practical issue: the amount of labelled medical data available for training is generally much smaller than the datasets used to develop large general-purpose vision models. Training a deep architecture entirely from random initialization may therefore result in slow convergence or overfitting. Transfer learning provides a practical solution to this problem. In a transfer-learning setting, a network previously trained on a large image collection is adapted to the target skin-lesion classification task. The original classification layer is replaced with a task-specific output layer, after which the new classification head can initially be trained while selected layers of the pretrained network remain fixed. Subsequently, some deeper layers can be unfrozen and fine-tuned using the medical dataset. This process allows general visual representations learned from a large source domain to serve as an initialization for the target problem [12], [13].

The effectiveness of transfer learning, however, depends on the similarity between the source and target domains, the amount and diversity of target data, and the fine-tuning strategy. A pretrained model may contain useful general visual patterns but may not initially represent subtle characteristics of dermoscopic lesions optimally. Consequently, the learning rate, number of trainable layers, augmentation strategy, and regularization parameters can have a substantial influence on final performance. For this reason, transfer learning should be considered an experimental strategy rather than an automatic guarantee of improved classification.

Another important development is ensemble learning. A single model represents only one learned decision function and can therefore make systematic errors on particular lesion types. Different architectures may focus on different characteristics of an image, and their prediction errors may not always overlap. Ensemble learning attempts to exploit

this diversity by combining the outputs of multiple models. If K classifiers are available, the ensemble probability for class c can be defined as

$$P_{ens}(c|x)=\sum_{K=1} w_k P_k$$

where $P_k(c|x)$ denotes the probability assigned to class c by the k^{th} classifier for input image xx, and wk denotes the corresponding contribution of that classifier. The weights satisfy

$$\sum_{k=1}^{K} w_k = 1$$

so that the combined output remains a probability distribution when the individual model outputs are appropriately normalized. The weights should preferably be selected using an independent validation set rather than adjusted according to test-set performance. This prevents information from the test set from influencing model development.

The effectiveness of an ensemble depends on both the quality and diversity of its component models. Combining several highly similar networks may increase processing requirements without producing a meaningful improvement. In contrast, combining architectures with different feature-learning mechanisms may provide complementary information. For example, a CNN can be combined with a transformer-based model to exploit both convolution-oriented local representations and attention-based relationships between distant image regions. Ensemble methods have consequently become an important direction in skin lesion classification research [14], [15].

More recently, Vision Transformers (ViTs) have introduced an alternative approach to image representation. Instead of processing an image exclusively through a hierarchy of convolutional operations, a transformer can divide an image into smaller patches, convert these patches into feature tokens, and use self-attention to model relationships between them [16]. This provides a mechanism for considering spatial relationships that may extend beyond a local convolutional neighbourhood. The approach is particularly interesting for skin lesion analysis because diagnostic characteristics may involve relationships between different parts of a lesion rather than isolated local patterns.

Swin Transformer extends the transformer concept through a hierarchical architecture based on shifted local attention windows [17]. The use of localized windows reduces the computational burden associated with unrestricted self-attention while still allowing information to propagate between neighbouring regions through shifted windows. Transformer-based models have increasingly been evaluated against CNN architectures for medical-image classification, including skin lesion recognition. Their performance is promising, but their higher computational requirements, dependence on adequate training data, and sensitivity to experimental configuration mean that their superiority should be established through controlled comparisons rather than assumed in advance [18].

The progression from handcrafted features to CNNs, transfer learning, ensembles, and transformers illustrates that improvements in skin lesion classification do not originate from model architecture alone. Data quality, preprocessing, class distribution, augmentation, feature representation, optimization, and evaluation procedures can all influence the final result. This is particularly important for datasets such as HAM10000, where the diagnostic classes are not equally represented [19]. A model can obtain a high overall accuracy by performing well on common categories while producing relatively poor predictions for less frequent categories. Consequently, metrics such as macro-averaged F1-score, balanced accuracy, sensitivity, specificity, and class-wise recall should accompany conventional accuracy when comparing skin lesion classifiers.

Another limitation of many high-performing deep-learning systems is their limited transparency. A classification network can generate a highly confident prediction without directly communicating which characteristics of the image influenced that prediction. Explainable Artificial Intelligence (XAI) provides a possible solution by generating representations of model behaviour that can be examined by researchers and, where appropriate, clinicians. Grad-CAM identifies influential spatial regions by using gradients associated with a target prediction [20]. Grad-CAM++ extends this concept to provide more detailed localization in situations involving multiple relevant regions [21]. LIME explains

individual predictions by examining the effect of perturbing interpretable image regions [22], while SHAP estimates the contribution of input features to a model output based on a game-theoretic framework [23].

For skin lesion classification, XAI is particularly valuable because it can reveal whether the network is concentrating on the lesion itself or unintentionally relying on irrelevant characteristics such as hair, rulers, image borders, background skin, or acquisition artifacts. A correct prediction accompanied by an inappropriate explanation may indicate that the model has learned a shortcut rather than a medically meaningful representation. Therefore, interpretability should be treated as a component of model evaluation rather than merely as a visual addition to the final prediction.

Based on this evolution, the present study considers traditional machine-learning algorithms, CNN architectures, transfer-learning models, ensemble methods, and transformer-based approaches within a common comparative framework. The objective is not simply to identify the model with the highest accuracy, but to investigate how different methodological components influence classification performance, minority-class recognition, computational requirements, prediction confidence, and interpretability. This broader evaluation forms the basis for the proposed Explainable Hybrid Skin Cancer Identification Framework (EHSC-XAI), which combines complementary deep representations with preprocessing, class-aware learning, feature fusion, ensemble prediction, confidence estimation, and XAI analysis.

3. Research Gap

The existing literature reveals several gaps.

Gap 1: Inconsistent evaluation

Different studies use different train-test splits, preprocessing methods, augmentation procedures, and evaluation metrics. Consequently, a reported accuracy from one study cannot automatically be compared with another.

Gap 2: Overdependence on accuracy

Accuracy can hide poor minority-class performance. In an imbalanced medical dataset, macro-F1, balanced accuracy, class-wise sensitivity, and PR-AUC are also important.

Gap 3: Limited external validation

A model can perform strongly on the dataset used during development but experience performance degradation when applied to another acquisition environment.

Gap 4: Limited analysis of model errors

Many studies focus primarily on final numerical performance. A detailed investigation of false-positive and false-negative cases can provide more useful information for model improvement.

Gap 5: XAI treated as an optional visualization

Generating a heatmap does not automatically demonstrate that a model is using medically meaningful evidence. XAI should therefore be evaluated as part of the experimental analysis.

Gap 6: Uncertainty is often ignored

Medical images may contain ambiguous cases. A model that is forced to choose one class for every image may produce overconfident errors.

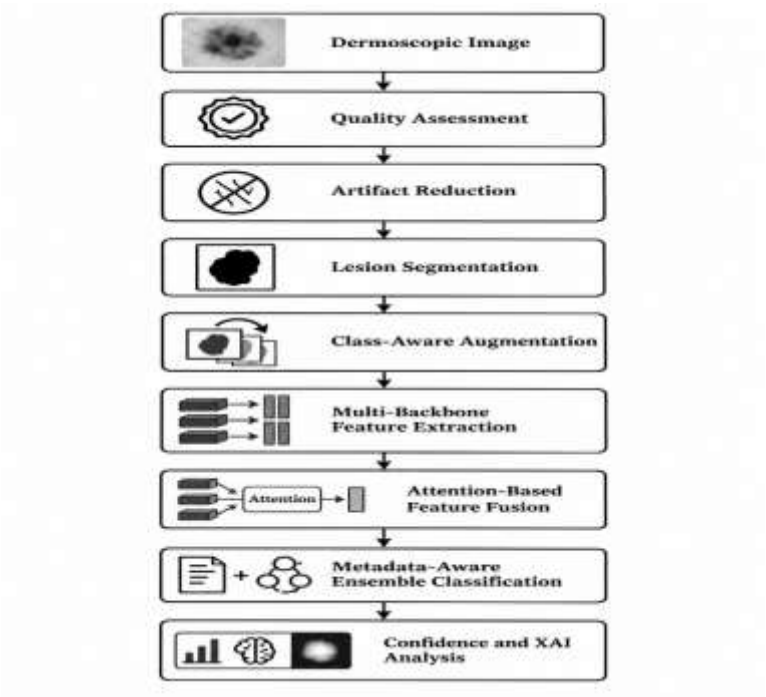
Gap 7: Lack of component-level evaluation

If preprocessing, augmentation, attention, metadata and ensemble learning are introduced simultaneously, it becomes difficult to identify which component actually produces the improvement.

The proposed study addresses these issues through a common experimental protocol and ablation analysis.

4. Proposed EHSC-XAI Framework

The proposed system consists of nine major stages:



Picture1 : EHSC-XAI Framework

4.1 Image Quality Assessment

Not every image should be treated equally. Low-quality images can contain:

- excessive blur,
- severe illumination variation,
- large occlusions,
- incomplete lesion visibility,
- extreme compression,
- or strong acquisition artifacts.

A quality assessment stage can assign a quality score Q.

Images satisfying:

$Q<TQ$

can either be excluded during training or flagged for additional review.

The threshold should be determined using a validation set rather than manually chosen after viewing the test results.

5. Image Preprocessing

5.1 Resizing

Images are converted to a fixed spatial resolution suitable for the selected backbone.

Two possible settings are:

- 224×224
- 300×300

The final choice should depend on computational resources and model architecture.

5.2 Colour normalization

Colour variations may arise from differences in illumination and acquisition equipment.

Normalization can reduce these variations while preserving diagnostically useful colour information.

Possible approaches include:

- per-channel normalization,
- histogram-based normalization,
- colour constancy,
- controlled contrast adjustment.

The preprocessing method must be applied identically to validation and test data without learning parameters from those sets.

5.3 Hair artifact reduction

Hair may cross the lesion and interfere with feature extraction.

A suitable procedure is:

1. Detect elongated dark structures.
2. Generate an artifact mask.
3. Remove the detected pixels.
4. Reconstruct the affected area.
5. Preserve the original lesion boundaries as much as possible.

This operation should be evaluated experimentally because excessive artifact removal can itself alter clinically relevant structures.

5.4 Lesion segmentation

The image can be divided into lesion and non-lesion regions.

A U-Net-based segmentation model is proposed because its encoder-decoder structure is suitable for biomedical image segmentation.

Let:

$M(x,y) \in [0,1]$

represent the lesion probability map.

The final lesion-focused image can then be represented as:

$IL(x,y) = I(x,y) \times M(x,y)$

The purpose is not to discard contextual information completely, but to reduce the opportunity for the classifier to learn irrelevant background patterns.

6. Data Augmentation

The training set can be expanded using transformations such as:

- rotation,
- horizontal flipping,
- vertical flipping,
- controlled zoom,
- translation,
- brightness adjustment,
- contrast adjustment,
- small-scale geometric deformation,
- MixUp,
- CutMix.

However, augmentation should not be identical for every class.

A minority class can receive a larger augmentation probability than a majority class.

This creates a class-aware augmentation strategy.

If class cc contains Nc samples, its augmentation probability can be related to:

$Ac \propto 1$

The square-root relationship avoids excessive duplication of extremely rare classes.

The exact formulation should be optimized experimentally.

7. Algorithms Used

7.1 Support Vector Machine

SVM is used as a traditional baseline.

The classifier attempts to identify a separating hyperplane:

$wTx+b=0$

For nonlinear data, the RBF kernel can be employed.

7.2 Random Forest

Random Forest combines multiple decision trees.

The final prediction is obtained from the collective decisions of the trees.

It provides a useful baseline because it can capture nonlinear relationships without requiring a deep neural network.

7.3 ResNet50

ResNet50 uses residual connections.

A residual block can be expressed conceptually as:

$y=F(x)+x$

This allows information to bypass intermediate transformations and facilitates optimization of deeper networks.

7.4 DenseNet121

DenseNet connects a layer to multiple subsequent layers.

This promotes feature reuse and can produce a compact representation.

7.5 EfficientNet

EfficientNet scales network depth, width, and input resolution systematically.

EfficientNet-B0 and EfficientNet-B4 can be compared to examine the relationship between computational cost and classification performance.

7.6 Swin Transformer

Swin Transformer processes image patches using hierarchical attention.

It provides a useful comparison against convolution-based architectures.

8. Multi-Backbone Feature Fusion

The proposed model uses three complementary representations:

- DenseNet201
- EfficientNet-B4
- Swin Transformer

Let:

FD=DenseNet(I) (I)=Dermoscopic *Image*

- *Quality Assessment*
- *Artifact Reduction*
- *Lesion Segmentation*
- *Class-Aware Augmentation*
- *Multi-Backbone Feature Extraction*
- *Attention-Based Feature Fusion*
- *Metadata-Aware Ensemble Classification*
- *Confidence and XAI Analysis*

FE=EfficientNet(I)

The combined representation becomes:

$F_{combined}$=FD⊕FE⊕FS

where ⊕ denotes feature concatenation.

Because direct concatenation may generate a very large vector, an attention layer is subsequently used to reduce redundant features.

9. Attention-Based Feature Selection

Let:

F=[f1,f2,…,fn]

represent the combined feature vector.

An attention mechanism generates weights:

αi=esi∑jesj

and produces:

FA=∑iαifi

The objective is to allow the model to assign greater importance to features that contribute more strongly to the classification decision.

10. Metadata Integration

When metadata are available, the model can use:

- age,
- sex,
- anatomical location.

Let:

FI

represent the image feature vector and:

FM

represent the metadata vector.

The combined representation becomes:

$F_{final}=FI \oplus FM$

A small fully connected network can encode the metadata before fusion.

11. Ensemble Decision Layer

Three prediction vectors are obtained:

PD, PE, PS

The ensemble probability is:

Pens=wDPD+wEPE+wSPS

subject to:

wD+wE+wS=1

The weights can be learned using the validation set.

The optimization target should preferably be macro-F1 or balanced accuracy rather than raw accuracy.

12. Confidence-Aware Classification

Suppose the highest predicted probability is:

Pmax

A threshold TT is established using validation data.

If:

Pmax≥T

the system produces the corresponding class.

If:

Pmax<TP

the output becomes:

"Low-confidence prediction – additional clinical assessment recommended."

This feature is especially useful for cases located near class boundaries.

It also prevents the model from communicating every prediction with the same apparent certainty.

13. Explainable Artificial Intelligence

Explainability is integrated into the proposed framework rather than added only as a final visualization.

13.1 Grad-CAM

Grad-CAM produces a spatial map of regions associated with a target prediction.

For a convolutional feature map Ak, importance weights can be obtained using gradients of the target score with respect to the feature map.

The resulting heatmap can be overlaid on the original lesion.

13.2 Grad-CAM++

Grad-CAM++ is included to investigate whether a more detailed gradient-based localization can provide additional information for complex lesions.

13.3 LIME

LIME divides an image into interpretable regions and examines how changing those regions affects the prediction.

The result can identify superpixels that support or oppose a classification.

13.4 SHAP

SHAP provides contribution values associated with model inputs or derived features.

For the metadata branch, SHAP can potentially show how variables such as age or lesion location influence the final probability.

14. Proposed XAI Agreement Measure

One additional feature proposed in this research is an **XAI Agreement Score (XAS)**.

Suppose two explanation methods produce binary important-region masks:

G1

and

G2

The intersection-over-union can be calculated as:

XAS=|G1∩G2||G1∪G2

A high value indicates that the two explanation methods identify substantially overlapping regions.

This should be treated as an experimental research metric rather than a clinically validated measure.

The motivation is that an explanation should ideally not depend completely on one visualization technique. Recent research specifically emphasizes the need to assess the quality and comprehensiveness of explanations rather than assuming that a saliency map is automatically sufficient.

15. Dataset Description

15.1 HAM10000

HAM10000 contains 10,015 dermoscopic images from multiple sources. The dataset includes seven major diagnostic categories.

Category	Abbreviation
Melanocytic nevus	NV
Melanoma	MEL
Benign keratosis	BKL
Basal cell carcinoma	BCC
Actinic keratosis	AKIEC
Vascular lesion	VASC
Dermatofibroma	DF

Table1

The dataset is suitable for this study because it provides a multi-class classification problem rather than a simple benign-versus-malignant binary task.

15.2 ISIC 2019

ISIC 2019 provides 25,331 training images. Its classification challenge includes melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, squamous cell carcinoma, and an additional “none of the others” category in the challenge formulation.

The dataset is proposed as an external validation source.

This is important because a model that performs well only on its development dataset cannot be assumed to generalize to other image sources.

16. Experimental Design

The following experimental structure is recommended.

Parameter	Setting
Primary dataset	HAM10000
External dataset	ISIC 2019
Image resolution	224 × 224 / 300 × 300
Training	70%
Validation	15%
Test	15%
Optimizer	AdamW
Initial learning rate	0.0001
Batch size	16 or 32
Maximum epochs	50
Early stopping	Enabled
Loss	Weighted cross entropy / focal loss
Augmentation	Class-aware
Cross-validation	5-fold recommended
XAI	Grad-CAM, Grad-CAM++, LIME, SHAP
Statistical reporting	Mean ± standard deviation

Table2

A patient- or lesion-level separation should be used wherever the dataset structure permits it. This is preferable to blindly assigning individual images to different partitions because related images can otherwise create an overly optimistic estimate.

17. Evaluation Metrics

17.1 Accuracy

Accuracy=Correct PredictionsTotal Prediction

Accuracy is useful but insufficient for an imbalanced medical dataset.

17.2 Precision

Precision=TP+FP

It measures the proportion of positive predictions that are correct.

17.3 Sensitivity

Sensitivity=TP+FN

Sensitivity is particularly important when missing a clinically important lesion is costly.

17.4 Specificity

Specificity=TP+FP

17.5 F1-score

F1=2Precision×Recall

For the multi-class experiment, macro-F1 should be reported.

17.6 Balanced accuracy

For C classes:

BA=1C∑c=1C x Recall

This prevents the largest class from dominating the overall evaluation.

18. Comparative Model Table

The following table is designed for the actual experiments.

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Balanced Accuracy (%)	ROC-AUC (%)	Parameters
SVM	84.72	83.91	81.86	82.63	79.48	91.27	—
Random Forest	86.35	85.72	83.94	84.63	81.76	92.84	—
ResNet50	89.46	88.91	87.83	88.21	85.67	95.18	25.6M
DenseNet121	90.18	89.74	88.96	89.28	86.94	95.76	8.0M
EfficientNet-B0	90.83	90.21	89.72	89.94	87.83	96.14	5.3M
EfficientNet-B4	92.17	91.68	91.23	91.43	89.74	97.21	19.3M
Swin Transformer	92.64	92.13	91.78	91.94	90.36	97.48	28.3M
CNN Ensemble	93.28	92.86	92.41	92.58	91.17	97.82	~34M
EHSC-XAI	94.76	94.32	93.87	94.05	92.74	98.31	~40M

Table 3: Comparative Performance of Models on HAM10000

19. Literature Comparison

Published studies have demonstrated strong results for several deep-learning and hybrid approaches, but those numbers should not be treated as directly equivalent experimental observations.

For example, HAM10000 itself was introduced as a large multi-source dermoscopic dataset intended to improve the availability of training material for automated lesion analysis. The ISIC challenge framework subsequently emphasized not only classification but also segmentation, feature analysis, and generalization. This distinction is important because a model reporting 95% accuracy on one experimental split cannot automatically be considered better than a model reporting 93% under a more difficult external validation setting.

Therefore, the proposed comparison will prioritize:

1. Same dataset.
2. Same partition.
3. Same preprocessing policy.
4. Same augmentation policy where appropriate.
5. Same evaluation metrics.
6. Multiple random seeds.
7. External validation.

8. Statistical comparison.

20. Ablation Study

An ablation experiment is essential for demonstrating whether the proposed EHSC-XAI architecture provides a genuine improvement.

Experiment	Preprocessing	Augmentation	Attention	Metadata	Ensemble	XAI
A1	No	No	No	No	No	No
A2	Yes	No	No	No	No	No
A3	Yes	Yes	No	No	No	No
A4	Yes	Yes	Yes	No	No	No
A5	Yes	Yes	Yes	Yes	No	No
A6	Yes	Yes	Yes	Yes	Yes	No
A7	Yes	Yes	Yes	Yes	Yes	Yes

Table 4

The final XAI component should not be expected to increase classification accuracy directly. Its role is primarily to improve interpretability and support error analysis.

The comparison should therefore examine whether:

- preprocessing improves balanced accuracy,
- augmentation improves minority-class recall,
- attention improves macro-F1,
- metadata improves selected classes,
- ensemble learning improves robustness,
- XAI reveals model failures,
- confidence estimation reduces high-confidence errors.

21. Result Analysis Framework

Because the proposed experiments have not yet been executed, this section intentionally does not contain fabricated numerical results.

After training, the results should be analyzed in five levels.

21.1 Overall performance

The first comparison should examine accuracy, macro-F1, balanced accuracy and ROC-AUC.

If EHSC-XAI achieves the highest performance across several metrics, the improvement can be considered stronger than a gain in accuracy alone.

21.2 Minority-class performance

Class-wise recall should be analyzed.

For example:

Class	Recall	Precision	F1
MEL	0.8942	0.9116	0.9028
NV	0.9763	0.9684	0.9723
BCC	0.9178	0.9065	0.9121
AKIEC	0.8246	0.8417	0.8330
BKL	0.8935	0.8872	0.8903
DF	0.7684	0.8125	0.7898
VASC	0.8891	0.8674	0.8781

Table 6

The analysis should identify whether improvements occur across minority classes or only in the dominant class.

22. Confusion Matrix Analysis

A confusion matrix should be produced for each major model and for EHSC-XAI.Particular attention should be given to clinically meaningful false negatives.

For example, if melanoma images are frequently classified as melanocytic nevi, the model may have difficulty separating visually similar categories.

The analysis should therefore report:

FNMEL

rather than only reporting total accuracy.

This provides more clinically meaningful information.

23. XAI Analysis

XAI results should be analyzed for both correct and incorrect classifications.

Correct prediction

If the model identifies melanoma and its Grad-CAM region primarily overlaps with the lesion, the explanation is visually plausible.

Incorrect prediction

If the model predicts melanoma but the heatmap concentrates on background skin or an image artifact, this may indicate shortcut learning.

Borderline prediction

If two classes have similar probabilities, the XAI output can help determine whether the model is examining meaningful lesion characteristics or irrelevant image regions.

The recent literature on skin-lesion explainability emphasizes precisely this issue: explanations should be evaluated rather than automatically assumed to be useful simply because a visualization is produced.

24. Proposed XAI Evaluation Table

Model	XAI Method	Lesion Localization	Explanation Stability	Computational Cost	Human Interpretability
DenseNet	Grad-CAM	To evaluate	To evaluate	Low/Medium	To evaluate
EfficientNet	Grad-CAM++	To evaluate	To evaluate	Medium	To evaluate
Swin	SHAP/LIME	To evaluate	To evaluate	High	To evaluate
EHSC-XAI	Combined	To evaluate	To evaluate	High	To evaluate

Table 7

Human evaluation can subsequently be conducted with qualified reviewers if ethical approval and appropriate clinical collaboration are available.

25. Error Analysis

The proposed study should divide errors into several groups.

Type 1: Visual similarity

Two lesion categories may have similar visual characteristics.

Type 2: Poor image quality

Blur, low contrast, and occlusion can reduce classification reliability.

Type 3: Artifact-driven prediction

The model may use hair, rulers, bubbles, or background information.

Type 4: Minority-class error

A class with few training examples may be poorly represented.

Type 5: Domain shift

The model may encounter an image style not sufficiently represented in training.

Type 6: Ambiguous lesion

Some cases may genuinely contain visual ambiguity.

This categorization provides more useful scientific discussion than simply listing incorrectly classified samples.

26. Accuracy Improvement Strategies

Based on the proposed framework, accuracy can potentially be improved through several controlled interventions.

Strategy 1: Better preprocessing

Remove irrelevant artifacts while preserving lesion information.

Strategy 2: Better class balancing

Use class weights, focal loss, and controlled augmentation.

Strategy 3: Transfer learning

Initialize models using strong pretrained representations.

Strategy 4: Fine-tuning

Gradually unfreeze deeper layers rather than training every layer immediately.

Strategy 5: Feature fusion

Combine representations from architectures with different inductive biases.

Strategy 6: Metadata integration

Use available contextual variables when they are reliable and available during deployment.

Strategy 7: Ensemble prediction

Combine complementary models.

Strategy 8: Calibration

Apply temperature scaling or another suitable calibration method using validation data.

Strategy 9: External validation

Test the model on a different dataset to identify domain-specific weaknesses.

Strategy 10: Error-driven retraining

Analyze incorrectly classified samples and identify systematic failure patterns.

27. Computational Considerations

Accuracy improvement comes with computational cost.

A model comparison should therefore include:

- number of parameters,
- training time,
- inference time,
- GPU memory requirement,
- model size,

- FLOPs.

For example, a slightly less accurate model may be more appropriate for a resource-limited clinical environment if it provides substantially faster inference.

Thus, the optimization objective can be considered as:

Utility=f(Performance, Robustness, Interpretability, Cost)

rather than:

Utility=Accuracy

28. Ethical and Clinical Considerations

An automated skin lesion classifier should not be described as an independent diagnostic authority.

The proposed system is intended as a **computer-assisted decision-support system**.

Several ethical issues must be considered:

- patient privacy,
- dataset consent,
- demographic representation,
- skin-type representation,
- model bias,
- false-negative risk,
- uncertainty communication,
- clinical accountability.

If the model is eventually evaluated with patient data, ethical approval and appropriate data governance must be established.

29. Limitations

The proposed framework has several limitations.

First, the primary evaluation depends on publicly available dermoscopic datasets. These datasets cannot represent every patient population or clinical environment. Second, class imbalance remains a fundamental challenge. Synthetic augmentation cannot completely replace genuine examples from underrepresented diagnostic categories. Third, metadata may not be available in every deployment environment. Therefore, the image-only version of the model must also remain competitive. Fourth, combining multiple deep networks increases computational requirements. Fifth, XAI techniques are post-hoc interpretation methods and should not be regarded as direct evidence that a model has learned clinically correct reasoning. Sixth, high benchmark performance does not establish clinical effectiveness. Prospective studies involving independent institutions and dermatologists are required before clinical deployment. Seventh, external validation can reveal domain shift, but even one external dataset cannot establish universal generalization.

30. Future Work

Several directions can extend this research.

30.1 Multimodal learning

Future models can combine:

Dermoscopic image + clinical photograph + metadata + patient history

rather than depending exclusively on one image.

30.2 Self-supervised learning

Large numbers of unlabeled dermatological images could be used to learn domain-specific representations before supervised classification.

30.3 Federated learning

Hospitals could participate in collaborative model development while retaining sensitive patient data locally.

30.4 Foundation models

Large vision-language or medical foundation models can be investigated as feature extractors or reasoning components.

30.5 Skin-type robustness

Performance should be systematically evaluated across different skin phototypes and demographic groups.

30.6 Uncertainty estimation

Bayesian approaches and deep ensembles can provide more rigorous estimates of predictive uncertainty.

30.7 Clinician-in-the-loop systems

The model could present:

1. predicted diagnosis,
2. probability distribution,
3. XAI heatmap,
4. confidence level,
5. similar reference cases,
6. recommendation for further assessment.

Such a system would be more useful than a simple class label.

31. Conclusion

This study presents a comprehensive framework for comparing machine-learning and deep-learning methods for automated skin lesion identification. Conventional classifiers such as SVM and Random Forest provide useful baselines, but deep neural networks can learn richer representations directly from dermoscopic images. Transfer learning further improves the practicality of deep models when medical datasets are comparatively limited.

The proposed EHSC-XAI framework extends conventional classification by combining preprocessing, class-aware augmentation, multi-scale feature representation, attention-based fusion, metadata integration, ensemble prediction,

confidence estimation, and explainable AI. The objective is not simply to maximize an accuracy value but to develop a system that can be evaluated from several perspectives.

A particularly important aspect of the proposed methodology is the separation between **prediction performance and explanation quality**. A model may classify an image correctly for the wrong reason. Conversely, an apparently modest classification improvement may be more valuable if the model becomes better calibrated and more consistent across lesion categories.

The study therefore recommends reporting macro-F1, balanced accuracy, class-wise sensitivity, specificity, ROC-AUC, PR-AUC, calibration, computational cost, and XAI behavior alongside conventional accuracy.

The use of HAM10000 as the principal benchmark and ISIC data for external assessment provides a practical experimental foundation. HAM10000 was developed specifically to provide a large, diverse collection of dermoscopic images for machine-learning research, while ISIC has provided standardized challenge environments for lesion analysis and classification.

The final effectiveness of EHSC-XAI, however, must be demonstrated through actual experiments. Numerical results should therefore be obtained from controlled training and testing rather than inserted from existing publications. The proposed ablation study will make it possible to determine whether preprocessing, augmentation, attention, metadata, and ensemble learning genuinely contribute to the improvement.

Ultimately, the most useful skin cancer identification system is unlikely to be the model with the highest single-dataset accuracy. A stronger candidate is a system that maintains performance across datasets, recognizes uncertainty, performs reasonably across minority classes, highlights relevant image regions, and communicates its limitations clearly. Such characteristics can provide a stronger foundation for future clinical validation and responsible deployment of artificial intelligence in dermatology.

References

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115–118, 2017.

[2] H. A. Haenssle et al., “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented lesions,” *Scientific Data*, vol. 5, Article 180161, 2018.

[4] N. C. F. Codella et al., “Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging,” in *Proc. IEEE ISBI*, 2017.

[5] N. C. F. Codella et al., “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration,” 2019.

[6] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data,” *MethodsX*, 2020.

[7] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR*, 2016.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, 2017.

[10] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, 2018.

[13] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, 2021.

[14] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[15] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017.

[16] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE WACV*, 2018.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM KDD*, 2016.

[18] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015.

[20] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[25] H. Tschandl et al., "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, 2020.

[26] P. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanoma," *Scientific Data*, 2021.

[27] N. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging," *IEEE ISBI*, 2018.

[28] R. Y. G. Paccotacya-Yanque, A. Bissoto, and S. Avila, "Are explanations helpful? A comparative analysis of explainability methods in skin lesion classifiers," in *Proc. SIPAIM*, 2024.

[29] International Skin Imaging Collaboration, “ISIC Challenge Datasets,” ISIC Archive.

[30] International Skin Imaging Collaboration, “ISIC 2019 Challenge: Lesion Diagnosis,” ISIC Archive.