

Advanced Skin Lesion Classification Using EfficientNetV2-S with a Proposed Segmentation and Hybrid CNN- Transformer Framework

M. Sutha

ME Communication System
Dept. of . Electronics and Communication Engineering
PET Engineering College, Valliyur, Tirunelveli

T. Agnes Ramena, M.E

Assistant professor
Electronics and Communication Engineering g
PET Engineering College, Valliyur, Tirunelveli

Abstract: Skin lesion analysis from dermoscopic images has become an important area of computer-aided medical research. Although deep learning models can learn useful visual patterns from these images, reliable classification remains difficult because different lesions can have similar appearances and clinically important classes may be underrepresented. Another challenge is that many image-classification systems process the complete image without explicitly separating the lesion from surrounding skin, hair, and other visual artefacts. In addition, predictions from complex models can be difficult to interpret. This study investigates an EfficientNetV2-S based approach for classifying eight categories of skin lesions and, separately, proposes a segmentation-guided hybrid framework that combines convolutional and transformer-based representations. The verified classifier was evaluated on 120 images and achieved an accuracy of 61.67%, balanced accuracy of 55.47%, macro-F1 of 53.28%, and macro ROC-AUC of 94.85%. The results were not uniform across the eight classes. Dermatofibroma and benign keratosis showed strong recall, while basal cell carcinoma and squamous cell carcinoma remained difficult to identify. The overall sensitivity for the malignant group was 37.5%, demonstrating that the verified classifier is not suitable for clinical diagnosis in its current form. To address these limitations, the proposed framework places lesion segmentation before classification and combines EfficientNetV2 features, which capture local visual patterns, with Vision Transformer features that provide broader contextual information. Grad-CAM is incorporated to provide visual explanations of model predictions. The proposed framework is presented as a research direction rather than as a validated clinical system, since controlled experiments comparing its individual components have not yet been completed. The study therefore establishes a verified baseline while defining a reproducible experimental pathway for evaluating segmentation, feature fusion, loss functions, test-time augmentation, and explainability in future work.

Keywords: Skin Lesion Classification, Dermoscopic Images, EfficientNetV2-S, Deep Learning, Vision Transformer, CNN-Transformer, U-Net++, TransUNet, Lesion Segmentation, Explainable AI, Grad-CAM, Computer-Aided Diagnosis

I. INTRODUCTION

Dermoscopic imaging provides a detailed view of skin lesions and is widely used to support the assessment of suspicious skin abnormalities. However, interpreting these images is not always straightforward. Some benign and malignant lesions share similar visual characteristics, while differences in colour, texture, shape, and surrounding structures can make classification difficult. The process also depends heavily on clinical experience, and access to dermatology specialists is not equally available in all settings. These challenges have encouraged researchers to investigate computer-aided methods that can assist with lesion analysis. Deep learning has become one of the most widely studied approaches for automated skin-lesion classification. Convolutional neural networks (CNNs), in particular, have demonstrated the ability to learn discriminative visual features directly from dermoscopic images. However, a conventional classifier generally receives the complete image as input. As a result, information such as surrounding skin, hair, lighting variations, and imaging artefacts may influence the learned representation even though these regions are not the primary target of diagnosis. Another limitation concerns the type of visual information captured by the model. CNNs are effective at learning local structures such as edges, textures, and fine-grained patterns, but their ability to represent long-range relationships is more limited. Vision Transformers (ViTs), in contrast, use self-attention to model relationships between different image regions and can therefore provide a broader representation of the image. This has motivated increasing interest in combining convolutional and transformer-based models rather than relying exclusively on

one architecture. Lesion segmentation provides another possible way to improve the classification process. By identifying the lesion region before classification, the model can concentrate more directly on the area containing clinically relevant information. Architectures such as U-Net++ and TransUNet have demonstrated the usefulness of encoder-decoder and transformer-assisted approaches for medical image segmentation. However, segmentation alone does not guarantee better classification. Its value needs to be evaluated experimentally alongside the classification model under the same data and evaluation conditions. Explainability is also important in medical image analysis. A prediction without any indication of which part of the image influenced the decision can be difficult for a clinician or researcher to inspect. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) can provide a visual representation of the regions contributing to a CNN prediction. Such explanations do not establish that a prediction is correct, but they can help researchers investigate whether the model is responding to the lesion or to irrelevant image features. This study therefore follows two related but distinct directions. First, an EfficientNetV2-S classifier is evaluated as a verified baseline for eight skin-lesion categories. The model is evaluated on 120 images, and its performance is reported using multiple classification metrics. Second, a segmentation-guided hybrid framework is proposed to address limitations associated with whole-image classification. The proposed design uses U-Net++ and TransUNet for lesion segmentation, EfficientNetV2 for local feature extraction, and a Vision Transformer for global contextual representation. The resulting features are intended to be combined before classification, with Grad-CAM used to provide visual explanations. An important distinction is maintained throughout this work: the EfficientNetV2-S results are experimentally verified, whereas the complete segmentation-guided hybrid framework remains a proposed architecture. No performance claim is made for the proposed components until controlled experiments are conducted. This distinction is particularly important because the verified baseline achieved only 37.5% sensitivity for the malignant group. Such a result makes clear that the current model should be treated as a research system rather than a clinical diagnostic tool. The study is therefore intended not only to report a baseline classification result but also to establish a structured framework for investigating whether lesion-focused processing, local-global feature fusion, and visual explanation can improve skin-lesion classification.

Research Objectives

The main objectives of this study are:

1. To evaluate an EfficientNetV2-S model for eight-class skin-lesion classification using a verified evaluation set.
2. To examine the strengths and weaknesses of whole-image classification, particularly for malignant lesion categories.
3. To investigate the role of lesion segmentation as a preprocessing stage before classification.
4. To design a hybrid CNN-Transformer architecture that combines local visual features from EfficientNetV2 with broader contextual features from a Vision Transformer.
5. To incorporate Grad-CAM-based visual explanations so that model predictions can be inspected spatially.
6. To establish a controlled ablation protocol for measuring the individual contribution of segmentation, feature fusion, loss functions, and test-time augmentation.
7. To develop a research-oriented computer-aided skin-lesion analysis workflow while clearly distinguishing experimental evidence from proposed functionality.

CONTRIBUTIONS

The contributions of this study are organized into two categories: experimentally verified components and proposed extensions for future evaluation.

Verified EfficientNetV2-S baseline: An EfficientNetV2-S classifier for eight skin-lesion categories was evaluated on 120 images. The evaluation reports overall accuracy, balanced accuracy, macro-F1, weighted-F1, ROC-AUC, and class-level precision, recall, specificity, and F1-score.

Class-imbalance handling support: The available training pipeline provides class-imbalance strategies, including weighted sampling, class-weighted cross-entropy, and optional weighted focal loss. These mechanisms are documented as available training functionality and are not presented as independently validated improvements in the reported experiment.

Cross-validation support: The training implementation provides stratified K-fold validation functionality, with five folds available as a default configuration. The use of this functionality for producing the evaluated checkpoint is not independently established.

Data augmentation support: The training pipeline provides image augmentation operations including horizontal flipping, vertical flipping, rotation, and colour perturbation. These operations are treated as available implementation capabilities rather than experimentally isolated contributions.

Test-time inference support: The implementation provides optional test-time augmentation and the deployed application supports multi-view inference. Because the exact inference configuration used to generate the reported evaluation metrics is not fully documented, no performance gain is attributed specifically to TTA.

Proposed segmentation-first framework: U-Net++ and TransUNet are proposed as alternative lesion-segmentation stages intended to focus subsequent classification on the lesion region. These components have not been experimentally evaluated in the present study.

Proposed CNN-Transformer fusion: A hybrid architecture combining EfficientNetV2 features with Vision Transformer features is proposed to combine local visual representations with broader contextual information. Its effectiveness remains to be established experimentally.

Proposed explainability component: Grad-CAM is proposed as an interpretability mechanism for examining the image regions contributing to model predictions. No quantitative or qualitative Grad-CAM evaluation is reported in the present study.

Research-oriented application integration: The verified classifier is integrated into a Flask-based skin-lesion analysis workflow containing authentication, image analysis, prediction, and reporting functionality. The application is intended for research and decision-support evaluation rather than autonomous medical diagnosis. Accordingly, the primary experimentally supported contribution of this study is the characterization of the verified EfficientNetV2-S baseline. Contributions 2–5 describe available implementation functionality, while contributions 6–8 represent proposed research extensions. The application integration described in contribution 9 has been verified at the workflow level but does not establish clinical effectiveness or diagnostic reliability.

II. LITERATURE REVIEW

Deep learning has become an important approach for automated skin-lesion analysis because dermoscopic images contain complex visual patterns that can be difficult to distinguish through conventional image-processing techniques. Differences in lesion colour, texture, shape, border structure, and surrounding skin can make automated classification challenging, particularly when visually similar lesions belong to different diagnostic categories. Recent research has therefore explored several complementary directions, including lesion segmentation, convolutional neural networks, Transformer-based models, hybrid architectures, and explainable artificial intelligence.

A. Segmentation-Guided Skin-Lesion Analysis

One important direction is the use of lesion segmentation before classification. Xie et al. [1] introduced a mutual bootstrapping deep convolutional neural network (MB-DCNN) that jointly considers segmentation and classification. In their approach, an initial segmentation network produces an approximate lesion region, which is then used to support localisation and classification. The classification information is subsequently used to improve segmentation. This interaction allows the model to make use of the lesion region rather than relying entirely on the complete dermoscopic image. The study reported Jaccard indices of 80.4% on ISIC-2017 and 89.4% on PH2, together with AUC values of 93.8% and 97.7%, respectively [1]. These results demonstrate the potential value of combining segmentation and classification rather than treating them as

completely independent tasks. More importantly, the work provides a basis for considering lesion localisation as part of the classification pipeline. However, the approach is primarily convolutional and does not explicitly incorporate Transformer-based global feature modelling or an explanation mechanism such as Grad-CAM.

B. Deep Learning for Lesion Segmentation

Accurate segmentation is particularly important when the classification system is expected to focus on the lesion itself. U-Net++ [3] extends the conventional encoder-decoder structure of U-Net through nested and dense skip connections. These connections are designed to improve the transfer of information between encoder and decoder feature maps and can help preserve fine boundary details. Such characteristics make U-Net++ relevant to dermoscopic images where lesion borders may be irregular, faint, or visually similar to surrounding skin. TransUNet [2] follows a different strategy by combining convolutional feature extraction with Transformer-based processing. CNN feature maps are converted into token representations and passed through a Transformer encoder, allowing the model to capture broader spatial relationships before reconstructing the segmentation output. This combination attempts to retain the local information provided by convolution while introducing the long-range modelling capability of self-attention. For the proposed framework, U-Net++ and TransUNet are considered as candidate segmentation architectures. Their inclusion is motivated by their complementary design characteristics: U-Net++ emphasises detailed feature propagation through nested connections, whereas TransUNet introduces global contextual modelling. However, these segmentation components remain proposed in the present study and have not yet been experimentally evaluated within the current implementation.

C. Efficient Convolutional Networks for Skin-Lesion Classification

Convolutional neural networks remain widely used for medical image classification because they are effective at learning local structures such as edges, textures, colour patterns, and shapes. EfficientNet [4] introduced a compound-scaling strategy that jointly considers network depth, width, and input resolution. Rather than increasing a single architectural dimension independently, the approach scales these factors in a coordinated manner. EfficientNet-B7, for example, was reported to achieve 84.3% ImageNet top-1 accuracy [4]. The EfficientNet family has also been explored specifically for skin-lesion analysis. Ma et al. [5] proposed EFFNet, which uses EfficientNetV2-derived features together with hierarchical bilinear pooling and a Random Forest classifier. Their reported results include 94.96% accuracy, 93.74% recall, 93.16% precision, and a 93.24% F1-score. Gessert et al. [6] investigated combinations of EfficientNet models operating at different image resolutions and incorporated patient metadata and augmentation into the classification process. These studies

demonstrate that the performance of an EfficientNet-based model depends not only on the backbone itself but also on the way features, image resolution, augmentation, and additional information are handled. This is relevant to the present work because the verified classifier uses EfficientNetV2-S as its backbone. At the same time, the relatively modest performance obtained by the verified model in this study should not be interpreted as a limitation of the EfficientNet family as a whole, since previous research has achieved substantially stronger results under different datasets and experimental settings.

D. Transformer-Based Skin-Lesion Classification

Although CNNs are effective at extracting local visual features, their conventional convolutional operations primarily build representations through local receptive fields. Transformer-based architectures provide an alternative by modelling relationships between image regions using self-attention. The Vision Transformer (ViT) introduced by Dosovitskiy et al. [7] represents an image as a sequence of patches and applies self-attention to model relationships between them. This allows information from distant image regions to contribute to the representation of a particular patch. Such global modelling can be useful for medical images in which diagnostic information may depend on relationships between different parts of a lesion. Hybrid CNN-Transformer approaches have consequently received attention in skin-lesion classification. Najjar et al. [8], for example, combined VGG19-based local feature extraction with Vision Transformer processing and used an augmentation strategy referred to as RSPDA. The study reported accuracies of 97.91% on MSK10000 and 97.10% on HAM10000 [8]. The relevance of these studies to the present work lies in the complementary nature of CNN and Transformer representations. CNNs can capture detailed local information such as texture and lesion structure, while Transformers can model broader relationships between image regions. The proposed framework therefore considers a hybrid EfficientNetV2 and ViT architecture rather than relying on either representation alone. This hybrid component, however, remains a proposed part of the current research and has not yet produced experimentally verified results.

E. Explainable Artificial Intelligence in Medical Image Analysis

High predictive performance alone is not sufficient for many medical imaging applications. A model should also provide information that helps users understand why a particular prediction was produced. Explainable artificial intelligence methods have therefore become increasingly relevant to computer-aided medical diagnosis. Grad-CAM [9] is one of the commonly used approaches for visualising the regions that contribute to a CNN prediction. It uses gradients associated with a target class to determine the importance of convolutional feature maps and produces a heatmap that can be overlaid on the original image. This provides a relatively

intuitive way of examining whether a model is focusing on the lesion region or on irrelevant image content. In the proposed framework, Grad-CAM is intended to complement the classification output by providing a visual indication of the image regions influencing the prediction. This is particularly relevant when the system is positioned as a research-oriented decision-support tool rather than as an autonomous diagnostic system. The Grad-CAM component is proposed and has not yet been experimentally validated in the current study.

F. Datasets Used in Skin-Lesion Research

The development of skin-lesion classification systems is closely related to the datasets used for training and evaluation. HAM10000 [10] contains 10,015 dermoscopic images and is widely used as a benchmark for pigmented-lesion classification. Its use in previous studies makes it useful for comparing the general direction of the proposed work with existing research. ISIC datasets are also important in this area. In particular, ISIC-2017 has been used for both segmentation and classification research, while ISIC-2018 provides a larger challenge-oriented setting involving lesion segmentation, attribute detection, and classification. PH2 has also been used in segmentation studies, including the MB-DCNN work discussed earlier. The available project pipeline can additionally accommodate datasets such as PH2 and Derm7pt. However, the exact training dataset used to produce the verified EfficientNetV2-S checkpoint is not established in the available project records. Therefore, the datasets discussed in this section are presented as research context rather than as claims about the origin of the verified checkpoint. This distinction is important because differences in dataset composition, class distribution, image quality, preprocessing, train-test splits, and evaluation protocols can substantially affect reported performance.

G. Overall Perspective from Previous Studies

The literature provides evidence for several useful design principles. Segmentation-based approaches demonstrate that identifying the lesion region can help direct subsequent analysis toward clinically relevant image content. EfficientNet-based studies show that carefully designed convolutional feature extractors can achieve strong performance when combined with suitable training strategies. Transformer-based research indicates that global relationships between image regions can complement the local representations learned by CNNs. Finally, explainability methods such as Grad-CAM provide a practical way to inspect the visual evidence behind a model prediction.

However, these ideas are often investigated separately or in different experimental settings. The existing literature therefore provides strong motivation for exploring an integrated pipeline in which lesion segmentation is followed by local and global feature extraction and an interpretable classification output.

H. Research Gap and Motivation

Despite the progress achieved by individual approaches, several limitations remain relevant to the development of a more comprehensive skin-lesion analysis system.

First, classification models that operate directly on the complete dermoscopic image may be influenced by surrounding skin, image artefacts, hair, or other non-lesion information. A segmentation-first strategy provides a possible way to reduce this influence by explicitly identifying the region of interest. Second, CNN-based and Transformer-based models capture different types of visual information. CNNs are well suited to local texture and structural patterns, whereas Transformers can model broader relationships across image regions. Combining these representations may therefore provide a more complete feature representation than relying on a single architecture. Third, segmentation and classification are frequently evaluated as separate tasks. A more integrated approach could investigate whether improvements in lesion localisation translate into improvements in downstream classification rather than assuming that better segmentation will automatically produce better diagnostic predictions. Fourth, explainability is not always incorporated directly into the classification workflow. For a medical-image analysis system, visual explanations can provide an additional layer of inspection and may help identify whether the model is responding to clinically meaningful regions. Finally, component-level evaluation is essential. A complex architecture should not be considered successful simply because it combines several advanced techniques. Each additional component should be evaluated against an appropriate baseline so that any improvement can be attributed to a specific design choice.

I. Positioning of the Present Study

Based on these observations, the present study is positioned as an investigation of a segmentation-first hybrid framework for skin-lesion classification. The proposed pipeline combines three main ideas: lesion-focused processing through segmentation, local and global representation learning through EfficientNetV2 and Vision Transformer features, and visual interpretation through Grad-CAM. The important point is that the proposed contribution does not claim that any individual component is new. U-Net++, TransUNet, EfficientNet-family networks, Vision Transformers, CNN-Transformer hybrids, and Grad-CAM have all been studied previously. The intended contribution is instead the systematic integration of these established techniques into a single research framework and the evaluation of whether their combination provides meaningful benefits over the verified EfficientNetV2-S baseline. At the current stage, only the EfficientNetV2-S classifier and its supporting evaluation and deployment workflow have been verified. The segmentation, CNN-Transformer fusion, and Grad-CAM components remain proposed extensions. Consequently, the literature supports the motivation and design of the proposed framework, but it

does not constitute experimental evidence that the proposed combination will outperform the verified baseline. This distinction is maintained throughout the study to avoid overstating the contribution.

III. PROBLEM STATEMENT AND RESEARCH GAP

A. Problem Statement

Automated skin-lesion classification is a challenging medical-image analysis problem because lesions belonging to different diagnostic categories can share similar visual characteristics. Variations in colour, texture, shape, border structure, lesion size, and surrounding skin can make reliable classification difficult even when high-quality dermoscopic images are available. A further challenge is that the complete dermoscopic image contains information that may not be relevant to the lesion itself. Hair, surrounding skin, image artefacts, and variations in acquisition conditions can influence the features learned by a classification model. When the lesion is not explicitly isolated, the network may learn visual patterns that are correlated with a particular class without necessarily representing meaningful lesion characteristics. Convolutional neural networks have demonstrated strong capability in extracting local image features such as texture, edges, and shape. However, a CNN-based model may have limited ability to represent relationships between distant regions of an image. Transformer-based architectures provide a complementary approach by modelling relationships between image patches through self-attention. At the same time, Transformer-only approaches can introduce higher computational requirements and may not always provide the local feature representation that is important for fine-grained lesion analysis. Another important consideration is interpretability. A classification probability alone does not explain why a model selected a particular lesion category. For a medical-image analysis system, the ability to inspect the image regions influencing a prediction can provide useful supporting information. Explainability does not guarantee that a prediction is correct, but it can help users examine whether the model is responding to a meaningful part of the image. The problem addressed in this study is therefore the development of a more focused and interpretable framework for automated skin-lesion classification. The proposed direction is to first identify the lesion region, then combine local and global visual representations for classification, and finally provide a visual explanation of the prediction.

B. Limitations of Existing Approaches

The literature reviewed in the previous section highlights several limitations that motivate this work. Many classification systems operate directly on the complete dermoscopic image without explicitly separating the lesion from its surroundings. This can make the classification process sensitive to background information and image artefacts. CNN-only architectures provide effective local feature extraction, but their representation of long-range

relationships is more limited than architectures explicitly designed around global attention. This creates an opportunity to complement convolutional features with Transformer-based representations. On the other hand, using a Transformer as the primary architecture does not automatically solve the problem. Vision Transformers can require substantial computational resources and may depend strongly on the amount and quality of training data. For skin-lesion classification, retaining detailed local texture and structural information remains important. A further limitation is the separation between segmentation, classification, and explainability. These components are often studied independently, while fewer systems attempt to integrate them into a single workflow in which lesion localisation supports classification and the final prediction is accompanied by an interpretable visual representation. Finally, generalisation remains an important concern. Results reported on one dataset cannot automatically be transferred to another because differences in patient populations, lesion distributions, image acquisition conditions, class imbalance, and evaluation protocols can significantly influence model performance.

C. Research Gap

Based on the existing literature, the main research gap is not the absence of individual techniques. Lesion segmentation has already been investigated, EfficientNet-based models have demonstrated strong classification capability, Vision Transformers have introduced global attention mechanisms, and Grad-CAM has provided practical visual explanations. The more specific gap lies in investigating how these complementary components can work together within a single skin-lesion analysis pipeline. The proposed framework addresses this gap by placing lesion segmentation before classification, allowing the downstream model to operate on a more focused representation of the dermoscopic image. EfficientNetV2 is intended to capture detailed local features, while a Vision Transformer is intended to model broader relationships between image regions. Their representations are then combined to form the basis of the final classification. Grad-CAM is incorporated as an additional interpretation layer so that the prediction can be examined visually. This approach is motivated by the assumption that local and global information may provide complementary evidence for lesion classification. However, this assumption must be tested experimentally rather than treated as an established result.

D. Position of the Present Study

The present study is deliberately structured around a verified baseline and a proposed extension. The verified component is an EfficientNetV2-S classifier designed for eight skin-lesion categories. Its checkpoint has been successfully loaded and evaluated on a 120-image evaluation set. The resulting performance provides the baseline against which future improvements can be assessed. The proposed extension introduces a segmentation-first architecture using

U-Net++ and TransUNet, followed by hybrid feature extraction using EfficientNetV2 and Vision Transformer components. Grad-CAM is then intended to provide visual interpretation of the classification output. This distinction is important because the proposed components have not yet been experimentally validated in the current study. Therefore, the framework should be understood as a research design motivated by the literature rather than as a demonstrated improvement over the baseline.

E. Research Motivation

The motivation for developing the proposed framework comes directly from the weaknesses observed in the verified baseline and from the limitations identified in previous research.

The current EfficientNetV2-S model provides a useful starting point, but its performance is uneven across the eight lesion categories. In particular, the reported malignant sensitivity is limited, demonstrating that overall classification accuracy alone is not sufficient to judge the usefulness of a skin-lesion classifier. The model therefore provides a realistic baseline rather than a clinically acceptable solution. The proposed framework is intended to address this limitation through a more structured analysis pipeline. Instead of asking a single classifier to interpret the complete image, the proposed design separates the problem into lesion localisation, feature extraction, feature fusion, classification, and visual explanation. The expected benefit is not simply a higher numerical accuracy. A meaningful improvement should also demonstrate stronger performance on clinically important classes, better balanced classification behaviour, and explanations that focus on relevant lesion regions. Computational requirements must also be considered because a more complex architecture is only useful if the additional cost is justified by measurable improvements.

F. Scope of the Proposed Work

The scope of the proposed research is limited to computer-aided analysis of dermoscopic skin-lesion images. The system is intended as a research and decision-support framework rather than an autonomous diagnostic system.

The proposed pipeline consists of five major stages:

Input Dermoscopic Image → Lesion Segmentation → Local and Global Feature Extraction → Hybrid Classification → Visual Explanation

The segmentation stage is intended to isolate the lesion from the surrounding image. The feature-extraction stage combines convolutional and Transformer-based representations. The classification stage produces the predicted lesion category, while the explainability stage provides a visual indication of the regions contributing to that prediction.

The proposed architecture will require controlled experimentation to determine whether each additional component provides a measurable benefit. In particular, segmentation quality, classification performance, class-

specific sensitivity, computational cost, and explanation quality should be evaluated independently rather than assuming that architectural complexity will automatically lead to better performance.

G. Expected Research Contribution

The intended contribution of this work is therefore a systematic investigation of a segmentation-first hybrid approach rather than the introduction of an entirely new neural-network architecture. The study establishes a verified EfficientNetV2-S baseline and provides a structured path for extending that baseline with lesion segmentation, CNN-Transformer feature fusion, and explainability. This provides a clear experimental foundation for determining whether the combination of these established techniques can improve skin-lesion classification.

Importantly, the current results do not yet demonstrate that the proposed framework closes the identified research gaps. That evidence can only be established after the proposed components are implemented, evaluated on controlled data, and compared directly with the verified baseline. Maintaining this distinction strengthens the scientific validity of the study and prevents the proposed architecture from being presented as experimentally proven before the required experiments have been completed.

IV. VERIFIED MODEL VERSUS PROPOSED FRAMEWORK

A clear distinction is maintained between the experimentally verified model and the proposed framework. The verified component is an EfficientNetV2-S classifier, while segmentation, CNN-Transformer fusion, and Grad-CAM are proposed extensions that require further experimental validation.

A. Verified Model

The verified model uses EfficientNetV2-S to classify dermoscopic images into eight lesion categories: actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevus, melanoma, squamous cell carcinoma, and vascular lesion. Evaluation on 120 images achieved 61.67% accuracy, 55.47% balanced accuracy, 53.28% macro-F1, and 94.85% macro ROC-AUC. Performance was uneven, with 37.5% malignant-class sensitivity, including 18.18% recall for basal cell carcinoma and no correctly identified squamous cell carcinoma samples. These results establish the EfficientNetV2-S model as the verified baseline.

B. Proposed Framework

The proposed framework extends this baseline using the following pipeline:

Dermoscopic Image → Preprocessing → Lesion Segmentation → EfficientNetV2 + Vision Transformer → Feature Fusion → Classification → Grad-CAM

U-Net++ and TransUNet are proposed for lesion segmentation, while EfficientNetV2 and Vision Transformer branches are intended to combine local and global image features. Grad-CAM is proposed to provide visual interpretation of model predictions. These components have not been experimentally validated in the current study. Therefore, they are presented as proposed extensions rather than reported results. Future experiments should evaluate whether each component improves macro-F1, balanced accuracy, and malignant-class sensitivity compared with the verified EfficientNetV2-S baseline.

V. PROPOSED SEGMENTATION-FIRST HYBRID FRAMEWORK

A. Overview

The proposed framework processes a dermoscopic image through lesion segmentation, hybrid feature extraction, classification, and visual explanation. U-Net++ and TransUNet are considered for segmentation, followed by EfficientNetV2 and Vision Transformer branches for feature extraction. Their features are then fused before classification, while Grad-CAM provides a visual representation of the regions contributing to the prediction.

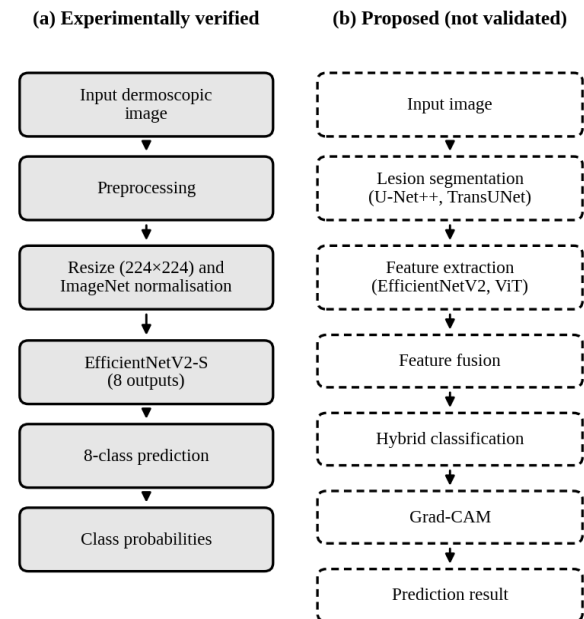


Fig. 1. (a) Experimentally verified workflow. (b) Proposed workflow; dashed boxes denote components that are not validated.

B. Proposed Preprocessing

The proposed preprocessing stage includes hair and artefact removal using black-hat filtering and inpainting, followed by

noise reduction, normalization, and image augmentation. These operations are intended to reduce visual artefacts that may interfere with lesion boundaries and classification. However, the implementation of black-hat filtering and inpainting in the training pipeline has not been established. The verified evaluation pipeline uses resizing, tensor conversion, and ImageNet normalization.

C. Proposed Segmentation Stage

Lesion segmentation is intended to focus the classifier on the relevant lesion region while reducing the influence of surrounding skin and image artefacts. U-Net++ [3] uses nested skip connections to preserve detailed boundary information, whereas TransUNet [2] combines CNN features with Transformer-based global context. The proposed segmentation loss combines Dice loss and binary cross-entropy. The relative performance of U-Net++ and TransUNet would need to be determined using quantitative segmentation metrics; such evaluation has not yet been performed.

D. Proposed Hybrid Classification

The proposed classifier combines EfficientNetV2 and Vision Transformer features. EfficientNetV2 is intended to capture local characteristics such as texture and shape, while the Vision Transformer models relationships between image patches [7]. The two feature representations are concatenated before classification:

$$[F_{\text{hybrid}}] = [F_{\text{CNN}} \parallel F_{\text{ViT}}]$$

followed by a softmax classification layer.

The exact layer configuration, embedding dimensions, attention heads, Transformer depth, and fusion dimensions have not yet been finalized. Therefore, these architectural choices remain part of the proposed experimental design.

E. Proposed Explainability

Grad-CAM [9] is proposed to generate class-specific heatmaps showing the image regions contributing to a prediction. Such visualizations could help determine whether the model focuses on the lesion rather than irrelevant structures such as image borders or artefacts. No quantitative or qualitative Grad-CAM evaluation has been performed in the current study.

F. Technical Background

The proposed framework is based on established techniques from the literature. EfficientNet uses compound scaling to balance network depth, width, and input resolution [4]. Vision Transformers divide images into patches and use self-attention to model relationships between them [7]. U-Net++ employs nested skip connections to improve feature propagation between encoder and decoder stages [3], while TransUNet incorporates Transformer-based global context into segmentation [2]. Grad-CAM uses class-specific gradients to generate spatial importance maps for visual

interpretation [9]. These mechanisms provide the technical basis for the proposed framework but do not represent experimentally obtained results from the present study.

G. Computational Considerations

The proposed framework introduces additional computational requirements through lesion segmentation, a second Vision Transformer feature-extraction branch, and Grad-CAM processing. Consequently, its effectiveness should be evaluated together with computational cost. Training time, parameter count, GPU memory usage, and inference time are not currently available for the verified model. These measurements should therefore be recorded for the baseline and proposed configurations before drawing conclusions about the practical benefit of the hybrid architecture.

VI. DATASET AND EXPERIMENTAL SETUP

A. Image Data

The verified EfficientNetV2-S model was evaluated on a set of 120 dermoscopic images covering eight skin-lesion classes: actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevus, melanoma, squamous cell carcinoma, and vascular lesion.

The dataset used to produce the checkpoint, its exact composition, and the data split are not currently established. Therefore, the reported evaluation should be interpreted as a verification of the available checkpoint rather than as a fully reproducible training experiment.

TABLE IV. CLASS DISTRIBUTION OF THE EVALUATED IMAGES

Class	Images	Category
Actinic Keratosis (AK)	16	Non-malignant*
Basal Cell Carcinoma (BCC)	11	Malignant
Benign Keratosis (BKL)	18	Non-malignant
Dermatofibroma (DF)	17	Non-malignant
Melanocytic Nevus (NV)	14	Non-malignant
Melanoma (MEL)	19	Malignant
Squamous Cell Carcinoma (SCC)	10	Malignant
Vascular Lesion (VASC)	15	Non-malignant
Total	120	40 malignant

The malignant group in this paper consists of BCC, MEL and SCC, which together account for the 40 malignant cases. The remaining five classes are grouped here as non-malignant for reporting purposes only; this is not a clinical statement about actinic keratosis.

B. Input Preprocessing

The verified evaluation pipeline resizes input images to 224 × 224 pixels, converts them to tensors, and applies ImageNet normalization before inference.

The training code additionally provides options for augmentation, including horizontal and vertical flipping, rotation, and colour variation. It also supports class-imbalance handling through weighted sampling, class-

weighted cross-entropy, and weighted focal loss. However, the exact configuration used to train the evaluated checkpoint has not been established.

C. Validation and Evaluation

The implementation provides support for stratified five-fold cross-validation and optional test-time augmentation. These capabilities are part of the available training pipeline, but their use in producing the verified checkpoint cannot be confirmed.

The verified model was evaluated using accuracy, balanced accuracy, macro-F1, weighted-F1, per-class precision, recall, specificity, and ROC-AUC. These measures provide complementary views of performance, particularly where class distributions and class-level performance are uneven.

D. Reproducibility Status

The checkpoint and evaluation results are available, but several details required for complete reproduction remain unavailable. These include the dataset used for checkpoint training, the exact train-test split, training configuration, hardware, exact package versions, and the prediction file or confusion matrix used for evaluation. Consequently, the current results provide a useful baseline for the proposed framework, but they should not be presented as a fully reproducible training study until these details are established.

E. Experimental Scope

Only the EfficientNetV2-S classifier is experimentally verified in this study. The proposed U-Net++, TransUNet, hybrid EfficientNetV2-Vision Transformer architecture, and Grad-CAM components are not included in the reported evaluation. Future experiments should use a fixed and documented dataset split and identical evaluation conditions when comparing the proposed components with the verified baseline.

VII. VERIFIED MODEL AND TRAINING CONFIGURATION

A. Verified Model

The experimentally verified classifier is an EfficientNetV2-S model with an eight-class output layer and an input resolution of 224×224 pixels. The available checkpoint has a size of 77.87 MiB and is identified by the model version skin-efficientnetv2s-research-2. Its metadata specifies a validation status of “research-only” and stores a temperature value of 0.5. The checkpoint was loaded using strict state-dictionary matching, and the loading process completed successfully. A forward pass using a single input image also completed successfully and produced an output tensor with the expected shape of (1, 8). These checks establish that the stored checkpoint is structurally compatible with the EfficientNetV2-S architecture used for inference. They do

not, by themselves, establish predictive quality; model performance is evaluated separately in Section IX.

TABLE VI. VERIFIED CHECKPOINT PROPERTIES

Property	Value
Architecture	EfficientNetV2-S
Output classes	8
Input size	224×224
Checkpoint size	77.87 MiB
model_version	skin-efficientnetv2s-research-2
validation_status	research-only
Temperature	0.5
Strict state-dict load	Successful
Forward pass	Successful; logits shape (1, 8)

B. Training Configuration Available from the Project Scripts

The project includes a training script and a v2 training workflow with configurable training options. However, these settings cannot be assumed to represent the configuration used to produce the verified checkpoint.

In particular, the default train.py configuration uses a timm EfficientNet-B3 with 384×384 inputs, whereas the verified checkpoint is an EfficientNetV2-S model with 224×224 inputs. The available configurations therefore describe possible training setups rather than the confirmed history of the evaluated checkpoint.

TABLE VII. TRAINING SCRIPT DEFAULTS AND v2 TRAINER OPTIONS

Item	train.py defaults	v2 trainer (README)
Backbone	efficientnet_b3 (timm)	EfficientNetV2-S (optional ensemble)
Epochs	20	30 (example command)
Learning rate	1e-4, AdamW	Warm-up + cosine; backbone at $0.1 \times$ head rate
Scheduler	ReduceLROnPlateau (max, patience 3, factor 0.5)	Warm-up + cosine
Loss	Weighted CE; optional focal ($\gamma = 2$)	Focal + weighted CE
Regularisation / stability	Optional AMP	Gradient clipping, accumulation, early stopping
Model selection	Best validation mean AUROC	Best validation macro-F1
Splits / folds	5 stratified folds	Lesion/patient-level GroupShuffleSplit
Other	Workers 4; seed 42; optional TTA; resume; ensemble evaluation	Optional ensemble: EfficientNetV2-S, ConvNeXt-Tiny, ViT-B/16, Swin-T

The v2 workflow also describes an optional ensemble in which temperature-scaled outputs from multiple networks are combined. The checkpoint evaluated in this study, however, is a **single EfficientNetV2-S model**, not an ensemble. Because the exact training configuration of the checkpoint is unavailable, the values in Table VII are reported only as documented project settings and should not

be interpreted as the confirmed training history of the evaluated model.

C. Checkpoint Metadata and Calibration

The checkpoint stores a temperature value of 0.5. In conventional temperature scaling, the logits z are divided by a temperature T before the softmax operation:

$$p = \text{Softmax}(z / T)$$

A value below one produces a sharper probability distribution, whereas a value above one produces a softer distribution. However, the available records do not establish whether this temperature value was applied when generating the probabilities used for the reported evaluation. Therefore, the stored temperature is treated as checkpoint metadata rather than as evidence of calibrated predictive probabilities. A synthetic verification input produced a top-class probability of 15.95%. This confirms that the inference pipeline can produce a probability distribution, but it is not a calibration result and is not part of the reported performance evaluation.

D. Application Inference and Test-Time Augmentation

The deployed application supports multi-view inference using image transformations that include flips, rotations, full-frame resizing, and a lesion-guided crop. The application can flag cases for “Urgent dermatology review” when predictions across supported views differ by the configured threshold.

However, the exact inference configuration used to produce the metrics reported in Section IX is not documented sufficiently to determine whether test-time augmentation was applied to those evaluation results. The reported metrics are therefore not attributed specifically to TTA. The multi-view inference functionality available in the application should also be distinguished from the optional TTA functionality provided by the training pipeline.

E. Scope of the Verified Experiment

Only the EfficientNetV2-S classifier is experimentally verified in this study. The proposed U-Net++, TransUNet, EfficientNetV2–Vision Transformer fusion, and Grad-CAM components are not included in the reported evaluation. The results should therefore be interpreted as a baseline characterization of the verified checkpoint rather than as an evaluation of the complete proposed framework.

VIII. MATHEMATICAL FORMULATION

This section presents the mathematical expressions associated with the proposed framework and the evaluation of the verified model. Eqs. (1)–(4) describe components of the proposed framework and have not been experimentally trained or evaluated in the present study. Eqs. (5)–(7) represent training and inference options available in the project implementation. Eqs. (8)–(14) define the evaluation measures used for the verified model. Eqs. (15)–(17) provide

the standard background formulations associated with the proposed components.

A. Proposed Hybrid Classifier and Losses

Let FCNN denote the feature representation obtained from the EfficientNetV2 branch and FViT denote the feature representation obtained from the Vision Transformer branch. The two feature representations are concatenated to form the fused representation H :

$$H = \text{Concat}(\text{FCNN}, \text{FViT}) \quad (1)$$

The fused representation is then passed through a fully connected classification layer followed by a softmax function to obtain the predicted class probabilities:

$$\hat{y} = \text{Softmax}(W \cdot H + b) \quad (2)$$

where W represents the weight matrix and b represents the bias vector of the final classification layer. The categorical cross-entropy loss for K classes is defined as:

$$L_{cls} = - \sum_i y_i \log(\hat{y}_i) \quad (3)$$

where y_i denotes the true class label and $\hat{y}_i$ denotes the predicted probability for class i .

For the proposed segmentation stage, the segmentation objective combines Dice loss with binary cross-entropy loss:

$$L_{seg} = L_{Dice} + L_{BCE} \quad (4)$$

The combined objective is intended to encourage accurate lesion-region overlap while maintaining pixel-level classification consistency. This segmentation formulation belongs to the proposed framework and was not experimentally evaluated in the present study.

B. Training Options in the Code

The training implementation provides mechanisms for addressing class imbalance through focal loss and class weighting. The focal-loss formulation reduces the contribution of easily classified samples while giving greater emphasis to difficult samples:

$$L_{focal} = (1 - p_t)^\gamma \cdot CE \quad (5)$$

where p_t represents the probability assigned to the true class, CE denotes the cross-entropy loss, and γ represents the focusing parameter. In the available training script, γ is set to 2. Class weighting is based on the inverse frequency of each class. For a class c containing n_c training samples, the class weight is expressed as:

$$w_c \propto 1 / n_c \quad (6)$$

where the proportionality indicates that the exact normalization of the weights is implementation-dependent.

The available checkpoint metadata records a temperature value of $T = 0.5$. Temperature scaling modifies the logits before the softmax operation as:

$$p = \text{Softmax}(z / T) \quad (7)$$

where z denotes the model logits and T denotes the temperature parameter. The available records do not establish whether temperature-scaled probabilities were used to calculate the metrics reported in Section IX.

C. Evaluation Measures

For each class, the numbers of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are determined using a one-versus-rest formulation. The corresponding evaluation measures are defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

$$\text{Precision} = TP / (TP + FP) \quad (9)$$

$$\text{Recall} = TP / (TP + FN) \quad (10)$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}) \quad (11)$$

$$\text{Specificity} = TN / (TN + FP) \quad (12)$$

For K classes, the macro-averaged F1-score is calculated as:

$$\text{Macro-F1} = (1/K) \sum_{k=1}^K F1_k \quad (13)$$

Similarly, balanced accuracy is calculated as the mean recall across all classes:

$$\text{Balanced accuracy} = (1/K) \sum_{k=1}^K \text{Recall}_k \quad (14)$$

where $K = 8$ represents the number of lesion classes.

Equation (8) represents accuracy in the one-versus-rest formulation. The overall multiclass accuracy reported in Section IX is calculated directly as the number of correctly classified images divided by the total number of evaluated images, resulting in 74 correct predictions out of 120 images.

Macro-F1 assigns equal importance to all classes and is therefore useful when class frequencies are uneven or when performance differs substantially across classes. Weighted-F1, in contrast, weights each class according to the number of evaluated samples. The reported ROC-AUC values are provided as macro and weighted averages across the eight classes. The available records do not specify the exact ROC-AUC implementation used to obtain these values.

D. Background Expressions for the Proposed Components

The following expressions provide the standard mathematical formulations associated with the proposed components. They describe the theoretical basis of the proposed framework and were not experimentally evaluated as independent components in the present study. For the EfficientNet family, the scaling of network depth, width, and

input resolution can be expressed using a compound scaling coefficient ϕ and corresponding scaling constants:

$$d = \alpha\phi, w = \beta\phi, r = \gamma\phi \quad (15)$$

where d represents network depth, w represents network width, r represents input resolution, and α , β , and γ are scaling constants. The self-attention mechanism used as the basis of the Vision Transformer branch is expressed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}(Q \cdot K^T / \sqrt{d_k}) \cdot V \quad (16)$$

where Q , K , and V represent the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the key vectors. For Grad-CAM, let A_k represent the k -th feature map of a convolutional layer, y_c represent the score associated with class c , and Z represent the number of spatial positions. The channel importance weight is calculated as:

$$\alpha_{kc} = (1/Z) \sum_i \partial y_c / \partial A_{ijk}$$

The corresponding Grad-CAM localization map is then expressed as:

$$\text{LGrad-CAM}_c = \text{ReLU}(\sum_k \alpha_{kc} A_k) \quad (17)$$

The Dice and Jaccard overlap measures provide segmentation-quality measures for a predicted lesion mask X and a reference mask Y . The Dice coefficient is defined as:

$$\text{Dice} = 2|X \cap Y| / (|X| + |Y|)$$

The Jaccard index is defined as:

$$\text{Jaccard} = |X \cap Y| / |X \cup Y|$$

These overlap measures would be used to evaluate the proposed segmentation components if the segmentation experiments are conducted. The Jaccard index is also the overlap measure reported by the MB-DCNN study cited as [1].

IX. RESULTS OF THE VERIFIED EFFICIENTNETV2-S MODEL

All results presented in this section correspond to the experimentally verified EfficientNetV2-S eight-class classifier evaluated on the 120-image evaluation set described in Section VI. The results apply only to the verified classifier. The proposed segmentation, hybrid CNN-Transformer, and Grad-CAM components were not experimentally evaluated and are therefore not included in the reported results.

A. Overall Performance

TABLE VIII. OVERALL PERFORMANCE (VERIFIED MODEL)

Metric	Value
Accuracy	61.67%
Balanced accuracy	55.47%
Macro-F1	53.28%

Metric	Value
Weighted-F1	59.30%
Macro ROC-AUC	94.85%
Weighted ROC-AUC	95.52%
Malignant cases	40
Malignant false negatives	25
Malignant sensitivity	37.5%

The model correctly classified **74 of the 120 images**, resulting in an accuracy of **61.67%**. Balanced accuracy and macro-F1 were lower, at **55.47%** and **53.28%**, respectively. This difference reflects the uneven performance across the eight classes. In contrast, the macro and weighted ROC-AUC values were **94.85%** and **95.52%**, indicating that the ranking-based discrimination measured by ROC-AUC was considerably more favourable than the final class-label accuracy

B. Per-Class Performance

TABLE IX. PER-CLASS METRICS (%)

Class	Recall	Specificity	Precision	F1
AK	87.50	89.42	56.00	68.29
BCC	18.18	88.99	14.29	16.00
BKL	94.44	94.12	73.91	82.93
DF	100.00	98.06	89.47	94.44
NV	28.57	100.00	100.00	44.44
MEL	68.42	100.00	100.00	81.25
SCC	0.00	99.09	0.00	0.00
VASC	46.67	86.67	33.33	38.89

The per-class results reveal a substantial variation in classification behaviour. **Dermatofibroma** achieved the strongest recall at **100%**, followed by **benign keratosis** at **94.44%** and **actinic keratosis** at **87.50%**. Melanoma achieved **68.42% recall**, with a precision and specificity of 100% in the evaluated set.

The weakest performance was observed for **squamous cell carcinoma**, for which none of the ten evaluated images was correctly identified. Its recall, precision, and F1-score were therefore all **0%**. Basal cell carcinoma also showed limited recognition, with **18.18% recall** and an F1-score of **16.00%**. Melanocytic nevus illustrates another important aspect of the results. Although its precision and specificity were both **100%**, its recall was only **28.57%**, meaning that the model produced no false-positive nevus predictions but correctly identified only four of the fourteen nevus images. Vascular lesion also showed relatively weak performance, with **46.67% recall** and the highest number of false positives among the eight classes.

TABLE X. PER-CLASS CONFUSION COUNTS (ONE-VERSUS-REST)

Class	TP	FN	FP	TN
AK	14	2	11	93
BCC	2	9	12	97
BKL	17	1	6	96
DF	17	0	2	101
NV	4	10	0	106
MEL	13	6	0	101

Class	TP	FN	FP	TN
SCC	0	10	1	109
VASC	7	8	14	91

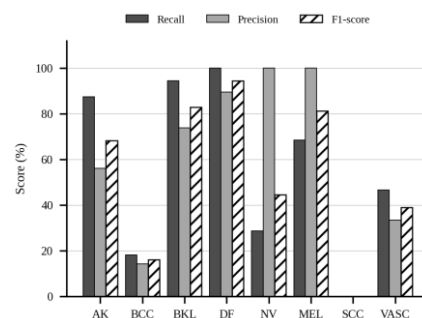


Fig. 2. Per-class recall, precision and F1-score of the verified model (values from Table IX).

Figure 2 provides a visual comparison of the per-class recall, precision, and F1-score values reported in Table IX. The distribution further demonstrates that the model does not perform uniformly across the eight categories.

As a consistency check, the false-positive counts across the eight one-versus-rest class evaluations sum to **46**. This agrees with the difference between the total number of evaluated images and the number classified correctly: **120 – 74 = 46**.

C. Malignant-Class Performance

For the analysis presented in this study, the malignant group comprises **basal cell carcinoma, melanoma, and squamous cell carcinoma**, giving a total of **40 images**.

Among these cases, **25 were missed at the corresponding exact class**, consisting of 9 BCC cases, 6 melanoma cases, and 10 SCC cases. The resulting exact-class malignant sensitivity is therefore **37.5% (15/40)**, where the 15 correctly identified cases are obtained from:

$$[2 + 13 + 0 = 15]$$

This value should not be interpreted as a strict binary benign-versus-malignant sensitivity. Such a binary metric could differ if a malignant image were incorrectly assigned to another malignant category. A complete binary malignant-versus-benign analysis would require the underlying multiclass confusion matrix, which is not available in the current evaluation record.

D. Verification Tests

In addition to the 120-image evaluation, a synthetic image was passed through the inference pipeline as a technical verification test. The model returned **Basal Cell Carcinoma** as the top predicted class with a probability of **15.95%**.

This test confirms that the inference path can execute from input image to model output. It is **not a clinical prediction, diagnostic result, or performance measurement**. For reference, an exactly uniform distribution over eight classes

would assign 12.5% to each class; therefore, the observed 15.95% probability is only modestly above that reference and should not be interpreted as evidence of model confidence.

The application-level verification also confirmed successful operation of the login, registration, login POST request, dashboard, analysis page, prediction insertion, and prediction page.

E. Additional Evaluation Information Not Available

Several items that would strengthen a complete experimental evaluation are not available in the current record. These include the original confusion-matrix figure, ROC curve figure, training and validation curves, model-comparison chart, training time, parameter count, GPU memory usage, inference time, and clinical prediction screenshots.

Their absence does not change the reported numerical results, but it limits the extent to which training behaviour, computational efficiency, and comparative model performance can be analysed.

F. Interpretation and Limitations of the Results

The results should be interpreted in the context of the available evaluation evidence. The evaluation contains only 120 images, with individual classes represented by 10 to 19 images. Consequently, a small number of additional correct or incorrect predictions can substantially change a class-level metric. The exact status of the evaluation set as a held-out test set is also not established. No external dataset was used, and the results represent a single checkpoint evaluated once, without repeated runs or fold-wise results from which performance variability could be estimated. Despite these limitations, the class-level pattern is clear. The model performs strongly for some classes but poorly for others, particularly the malignant categories. The 0% recall for SCC across ten evaluated images and the 18.18% recall for BCC indicate substantial difficulty in recognising these classes in the available evaluation set. The malignant sensitivity of 37.5% further highlights this limitation. These findings are therefore better interpreted as a diagnostic assessment of the current research checkpoint than as evidence of clinical readiness.

G. Summary of Key Findings

The verified EfficientNetV2-S model produced the following main results: 74 of 120 images were correctly classified, corresponding to 61.67% accuracy. Balanced accuracy and macro-F1 were 55.47% and 53.28%, respectively, reflecting uneven class-level performance. Macro and weighted ROC-AUC were 94.85% and 95.52%, respectively. Dermatofibroma, benign keratosis, and actinic keratosis achieved the highest recall values of 100%, 94.44%, and 87.50%, respectively. Melanoma achieved 68.42% recall with 100% precision and 100% specificity in the evaluated set.

Basal cell carcinoma and squamous cell carcinoma were the weakest classes, with recall values of 18.18% and 0%, respectively. The malignant group contained 40 images, of which 25 were missed at their corresponding malignant class, resulting in an exact-class malignant sensitivity of 37.5%. Overall, the verified results establish a measurable baseline for the subsequent discussion and for future evaluation of the proposed segmentation-first hybrid framework. They also show why additional validation is necessary before the proposed extensions can be considered improvements over the current model.

X. DISCUSSION

A. Strengths and Weaknesses

The verified EfficientNetV2-S model showed substantial variation in performance across the eight lesion classes. The strongest results were obtained for dermatofibroma, benign keratosis, and melanoma. Dermatofibroma achieved 100% recall and 94.44% F1-score, while benign keratosis achieved 94.44% recall and 82.93% F1-score. Melanoma achieved 68.42% recall, 100% precision, and 100% specificity in the evaluated set. The melanoma result should nevertheless be interpreted cautiously. Six of the 19 melanoma images were not correctly identified, meaning that a substantial proportion of melanoma cases were still missed. Therefore, the result does not support any claim of clinical suitability.

The weakest performance occurred for basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). BCC achieved only 18.18% recall and 16.00% F1-score, while SCC achieved 0% recall and 0% F1-score. The available evidence does not establish why these classes performed poorly. The training-data distribution, original confusion matrix, and controlled ablation results are unavailable. Possible explanations such as limited training examples or visual similarity between lesion categories should therefore be treated only as hypotheses.

B. Malignant-Class Sensitivity

The malignant-class analysis is one of the most important findings of the study. Of the 40 malignant images, only 15 were correctly identified at their corresponding malignant class, resulting in an exact-class malignant sensitivity of 37.5%. This means that 25 of the 40 malignant cases were missed in the evaluated set. Such performance is insufficient for clinical screening and supports the classification of the current system as a research-only model rather than a clinically validated diagnostic tool. The application documentation also requires dermatologist review of predictions and does not support assigning disease stage from an image. These restrictions are consistent with the limitations observed in the experimental results.

C. Accuracy and ROC-AUC

A notable characteristic of the results is the difference between classification accuracy and ROC-AUC. The model achieved 61.67% accuracy, while macro and weighted ROC-

AUC reached 94.85% and 95.52%, respectively. These measures capture different aspects of model behaviour. Accuracy evaluates the final predicted class, whereas ROC-AUC evaluates the ranking of positive and negative examples across decision thresholds. Therefore, a high AUC does not necessarily imply that the model produces accurate final class labels. One possible interpretation is that the model's output scores may contain useful class-ranking information that is not fully reflected in the final argmax decision. Calibration or class-specific decision thresholds could potentially affect this relationship. However, this possibility cannot be tested from the available evidence because the per-image probability outputs and the exact ROC-AUC computation procedure are not available. The high AUC values should therefore not be interpreted as evidence that the model is clinically performing at an equivalent level.

D. Statistical Uncertainty

The evaluation set is relatively small, with individual classes containing only 10 to 19 images. Consequently, a small change in the number of correct predictions can produce a noticeable change in a class-level percentage.

Using the reported counts, a 95% Wilson score interval for the overall accuracy of 74/120 is approximately 52.7%–69.9%. For the exact-class malignant sensitivity of 15/40, the corresponding interval is approximately 24.2%–53.0%.

These intervals are calculated from the reported counts and assume that the evaluated images can be treated as independent observations. They do not account for possible dependence between images originating from the same lesion or patient. They are therefore provided only as an indication of the uncertainty surrounding the reported proportions rather than as a complete statistical analysis.

E. Relation to Published Studies

The results reported in the literature review were obtained using different datasets, preprocessing procedures, model configurations, and evaluation protocols. Consequently, the performance values reported in Table I should not be directly compared with the 61.67% accuracy obtained by the present verified checkpoint.

The purpose of the literature comparison is to establish the broader research context rather than to claim superiority over previous methods. The difference between published results and the current baseline highlights the need for further experimentation, but the available evidence does not support attributing that difference to any particular factor because the dataset, split, and original training configuration of the verified checkpoint are not known.

F. Status of the Proposed Components

The project implementation provides support for several additional techniques, including weighted cross-entropy, focal loss, test-time augmentation, and model ensembling. However, these options have not been evaluated through controlled ablation experiments for the reported checkpoint.

Similarly, the proposed U-Net++, TransUNet, EfficientNetV2-Vision Transformer fusion, and Grad-CAM components have not produced experimental results in the current study. Therefore, no claim can be made that any of these components improves classification performance.

The motivation for the proposed segmentation-first hybrid framework is instead based on established approaches reported in the literature [1]–[3], [7]–[9]. Whether these components can improve the weak classes observed in the current baseline, particularly BCC and SCC, remains to be experimentally determined.

G. Class-by-Class Interpretation

Dermatofibroma and benign keratosis. These classes produced the strongest overall results. Dermatofibroma achieved 17/17 correct detections, while benign keratosis achieved 17/18. Their F1-scores of 94.44% and 82.93%, respectively, indicate that the model was able to separate these categories relatively well within the available evaluation set.

Actinic keratosis. Actinic keratosis achieved a relatively high recall of 87.50%, but its precision was considerably lower at 56.00%. The model produced 11 false-positive predictions for this class, indicating a tendency to assign the actinic keratosis label to images belonging to other categories.

Melanocytic nevus. The model achieved 100% precision and specificity for melanocytic nevus, but its recall was only 28.57%. None of the images from other classes was incorrectly identified as nevus, yet only four of the 14 nevus images were correctly detected. The reason for this imbalance cannot be established from the available training and prediction records.

Vascular lesion. Vascular lesion achieved 46.67% recall and 33.33% precision. It produced 14 false positives, the largest false-positive count among the eight classes, while eight of its 15 images were missed. This combination suggests difficulty in separating vascular lesions from other categories within the evaluated set.

Basal cell carcinoma and squamous cell carcinoma. These two malignant classes represent the most serious weaknesses of the verified model. Together they account for 19 of the 25 malignant false negatives. BCC produced only two correct predictions from 11 images, while SCC produced no correct predictions from its ten images. In particular, the zero-recall result for SCC indicates that this class requires substantial attention in any future model development and evaluation.

H. Implications for the Proposed Framework

The observed weaknesses provide a practical motivation for evaluating the proposed segmentation-first hybrid framework, but they do not prove that the proposed components will solve the identified problems.

If irrelevant background structures or image artefacts influence classification, lesion-focused segmentation may help the classifier concentrate on the relevant region. If classification errors are partly related to the need for broader

spatial relationships, the proposed Vision Transformer branch may provide complementary global information to the CNN features. Similarly, Grad-CAM could help visualise the regions contributing to individual predictions and provide a means of inspecting whether the model focuses on clinically relevant image regions. These are proposed hypotheses rather than findings demonstrated by the present experiments. The planned evaluations described later in the paper are therefore necessary to determine whether segmentation, hybrid feature fusion, and visual explanation provide measurable benefits over the verified EfficientNetV2-S baseline.

Overall, the discussion supports a clear conclusion: the current EfficientNetV2-S checkpoint provides a measurable research baseline, but its uneven class performance and particularly low malignant sensitivity prevent it from being considered clinically reliable. The proposed framework should therefore be evaluated as a research extension of the baseline, with improvements demonstrated through controlled experiments rather than assumed from the architecture alone.

I. Objectives Against Evidence

TABLE XI. PROJECT OBJECTIVES AND CURRENT EVIDENCE

Objective	Evidence at present
1. Improve classification with deep learning	Baseline established (61.67% accuracy); improvement not demonstrated
2. Investigate segmentation before classification	Proposed; not validated
3. Focus on the lesion region	Proposed; application README mentions lesion-guided crop in TTA, not evaluated
4. Combine local and global features	Proposed (EfficientNetV2 + ViT); no results
5. Incorporate Grad-CAM	Proposed; heatmap viewer mentioned in README; no evaluation
6. Integrated analysis system	Flask workflow implemented; application checks passed; no clinical validation

J. Application Inference Versus Reported Metrics

The deployed application uses a more extensive inference workflow than a single model forward pass. It supports test-time augmentation across multiple image views, including flips, rotations, full-frame resizing, and a lesion-guided crop, and it can flag cases when predictions across views differ by the configured threshold.

However, the evaluation records do not establish whether the metrics reported in Section IX were obtained using this multi-view application workflow, the four-view TTA option available in the training script, or a single-view inference procedure. The reported results should therefore be understood as measurements of the verified classifier, rather than as a confirmed evaluation of the complete deployed inference pipeline.

As a result, the behaviour of the full application on the same 120 images may differ from the reported metrics. The direction and magnitude of any such difference cannot be determined from the available evidence. A controlled comparison of single-view inference with the application's

multi-view TTA procedure is therefore identified as a planned experiment (E6 in Table XII).

XI. DEPLOYMENT AND PRACTICAL USE

The verified EfficientNetV2-S model has been integrated into a Flask-based web application that provides an interface for image analysis and prediction management. The application supports user registration and authentication, a dashboard, image analysis, prediction viewing, patient workflows, doctor and administrator review, and PDF report generation.

Three application roles are defined: patient, doctor, and administrator. Prediction and review records are stored in CSV files. The model is loaded lazily when the first image is submitted rather than during application startup, and uploaded files are subject to a 16 MB size limit.

A. Security Measures

The application includes several application-level security mechanisms, including CSRF protection, request-rate limiting, Talisman security headers with a configured content security policy, and audit logging. The documented default rate limits include 200 requests per day and 50 requests per hour, with a stricter limit of five requests per minute for login attempts. These measures are intended to improve the security and operational robustness of the web application. They should not, however, be interpreted as evidence that the underlying classification model is accurate, clinically reliable, or secure from a medical decision-making perspective.

B. Containerisation and Execution

The application provides a Docker-based deployment configuration using python:3.11-slim as the base image. The container runs the Flask application through Gunicorn with two workers on port 5000 and uses a non-root user.

The Docker Compose configuration provides environment variables including SECRET_KEY and FLASK_ENV=production and mounts persistent volumes for the CSV database, uploaded files, generated reports, and application logs. A Windows run.bat launcher is also provided for local execution.

These configurations support reproducible application deployment, although the available project information does not establish production-scale performance or operational capacity.

C. Continual-Learning Log

The application records reviews approved by doctors or administrators in a continual-learning log. The recorded status is approved_prototype_memory, while the associated model version is marked pending_validation.

Importantly, these reviewed cases are not automatically used to retrain the classifier. No automatic retraining schedule is currently defined. The log therefore functions as a record of reviewed cases and potential future training data rather than as an active continual-learning loop.

D. Operational Guidance

The project documentation recommends running the application behind a TLS reverse proxy, maintaining encrypted backups of persistent volumes, and requiring dermatologist review of predictions. It also explicitly advises against assigning an AJCC stage from an image.

The application additionally provides an “Urgent dermatology review” flag when predictions across supported inference views differ by the configured threshold. This mechanism should be regarded as a workflow safeguard rather than a clinically validated risk score.

Several operational characteristics have not yet been established, including measured inference time, a defined retraining schedule, and production monitoring metrics. These should be specified and evaluated before considering a production-oriented deployment.

E. Application Workflow

During normal use, a user registers and authenticates before accessing the relevant application workflow. An image can then be uploaded through the analysis interface, processed by the available model, and associated with a prediction record that can subsequently be viewed through the prediction interface.

Patient workflows are separated from doctor and administrator review functions. Approved reviews are recorded in the continual-learning log, and PDF reports can be generated from the available results.

The model is loaded when the first image is submitted rather than when the application starts. This lazy-loading strategy reduces the initial application startup requirement. If a validated checkpoint cannot be loaded, the application returns an “Unknown Lesion” output rather than presenting an unsupported class prediction.

F. Intended Use and Human Oversight

The application is intended for research and decision-support research, not for autonomous medical diagnosis. This distinction is supported by several design and documentation choices. The checkpoint is explicitly marked research-only, dermatologist review is required, and the application does not attempt to assign disease stage from an image. The application can also flag cases for “Urgent dermatology review” when predictions across inference views disagree by the configured threshold. While this may help identify potentially unstable predictions, the threshold itself has not been clinically validated in the present study. It therefore cannot compensate for systematic classification errors, particularly the poor malignant-class sensitivity observed in the verified evaluation. Human review remains essential because the current model achieved only 37.5% exact-class malignant sensitivity on the evaluated malignant cases. The application should therefore be considered an interface for research evaluation and decision support rather than a replacement for clinical assessment.

G. Maintenance and Long-Term Considerations

Several issues require attention before the application could be considered for broader or sustained deployment. First, application records are stored in CSV files. This approach is simple and transparent for a prototype, but its suitability for larger-scale workflows involving sensitive patient information requires careful assessment. Second, the continual-learning log records reviewed cases with the model version pending validation. These records have not been incorporated into an updated model, and no automated retraining schedule has been defined. Any future retraining should therefore use a controlled dataset and an explicit validation procedure rather than automatically updating the deployed checkpoint from reviewed cases. Third, production monitoring metrics have not yet been defined. A mature deployment would require systematic monitoring of operational behaviour and model performance so that degradation could be identified after deployment. Finally, inference latency, resource consumption, model-update procedures, data-retention practices, and production monitoring remain areas for future engineering work. These considerations are particularly important because the current system is a research prototype and the verified classifier has not demonstrated clinical reliability.

XII. LIMITATIONS, ETHICS AND PRIVACY

A. Limitations

The verified EfficientNetV2-S checkpoint is intended for research use and has not undergone clinical validation. Its overall accuracy was 61.67%, while exact-class malignant sensitivity was 37.5% for the malignant group defined in this study as basal cell carcinoma, melanoma, and squamous cell carcinoma.

The evaluation set contains only 120 images, with individual classes represented by 10–19 images. Consequently, individual predictions can have a noticeable effect on class-level metrics. The original dataset split and complete training history associated with the evaluated checkpoint are also not fully documented, which limits reproducibility and generalisation claims.

The study evaluates a single verified checkpoint without external-dataset validation or repeated independent experimental runs. The proposed U-Net++, TransUNet, CNN-Transformer fusion, and Grad-CAM components have not yet been experimentally validated. Their expected benefits should therefore be regarded as research hypotheses rather than established improvements.

These limitations define the scope of the present work as a verified baseline and research framework rather than a clinically validated diagnostic system.

B. Ethics and Privacy

The developed application includes application-level safeguards such as authentication, CSRF protection, rate limiting, security headers, audit logging, and role-based workflows. These mechanisms provide software-level

protection but do not establish clinical safety, diagnostic reliability, or regulatory compliance.

Model outputs should therefore be treated strictly as research and decision-support information. They should not be used for self-diagnosis, treatment selection, or to delay consultation with a qualified medical professional. In particular, the observed malignant-class errors demonstrate that a reassuring prediction cannot be interpreted as evidence that a lesion is benign.

The present study does not claim clinical deployment or regulatory validation. The available project records do not document a clinical ethics approval, informed-consent protocol, or a validated anonymisation procedure for identifiable patient images. Any future study involving such data should establish appropriate institutional ethical approval, informed consent where required, anonymisation or pseudonymisation, controlled access, secure storage, retention, and deletion procedures before data collection or deployment.

C. Threats to Validity

The principal threats to validity arise from the limited evaluation size, incomplete documentation of the checkpoint's original data split and training history, absence of external validation, and lack of experimental evaluation of the proposed framework.

The malignant sensitivity reported in this paper is specifically based on exact-class identification of basal cell carcinoma, melanoma, and squamous cell carcinoma. Alternative definitions of malignant detection, such as binary malignant-versus-benign classification, may produce different values.

Accordingly, the reported results should be interpreted as an experimentally verified baseline for subsequent development rather than evidence of clinical effectiveness.

XIII. PROPOSED EVALUATION PROTOCOL AND REPRODUCIBILITY

This section establishes a reproducible evaluation framework for the next stage of the study and documents the reproducibility status of the currently verified EfficientNetV2-S baseline. The proposed segmentation and CNN-Transformer components have not yet been experimentally evaluated; therefore, they are treated as planned research configurations rather than demonstrated improvements.

A. Controlled Experimental Design

The verified EfficientNetV2-S classifier is used as the reference configuration (E0). Subsequent experiments should isolate the contribution of segmentation, transformer-based feature extraction, feature fusion, loss-function design, and test-time augmentation. To ensure meaningful comparison, all primary experiments should use the same evaluation dataset, class definitions, preprocessing policy, and reporting metrics.

The planned configurations are:

- E0: EfficientNetV2-S using the complete input image experimentally verified baseline.
- E1: U-Net++ lesion segmentation followed by EfficientNetV2-S classification planned.
- E2: TransUNet lesion segmentation followed by EfficientNetV2-S classification planned.
- E3: EfficientNetV2-S and Vision Transformer feature fusion using the complete image planned.
- E4: Segmentation-guided EfficientNetV2-S and Vision Transformer feature fusion planned full framework.
- E5: Comparison of standard cross-entropy, weighted cross-entropy, and weighted focal loss planned.
- E6: Comparison of inference with and without test-time augmentation planned.
- E7: Grad-CAM analysis of the proposed model outputs planned explainability experiment.

The primary architectural comparison should be between E0 and E4. Configurations E1–E3 are intended to determine whether any observed change can be attributed to an individual architectural component. E5 and E6 evaluate training and inference strategies independently of the principal architectural design, while E7 evaluates visual interpretability.

B. Reproducibility Status of the Verified Baseline

The evaluated checkpoint and its performance measurements are available and provide a reproducible inference baseline. The verified checkpoint corresponds to an eight-class EfficientNetV2-S classifier with a 224×224 input resolution. Its stored size is approximately 77.87 MiB, and its model version is identified as skin-efficientnetv2s-research-2. Strict state-dictionary loading was successfully verified, and a forward pass produced the expected eight-class output.

The reported evaluation includes accuracy, balanced accuracy, macro-F1, weighted-F1, per-class precision, recall, specificity, and ROC-AUC. The class-wise evaluation counts and the reported performance values are also available. These records allow the reported inference-level results to be inspected and used as the reference point for subsequent controlled experiments.

However, complete reproduction of the original checkpoint-training process is not currently possible because several historical training records are unavailable. Specifically, the following information has not been established for the evaluated checkpoint:

- original training dataset identifier and exact dataset version;
- exact training, validation, and test partitions;
- confirmed training hyperparameters;
- confirmed preprocessing and augmentation configuration;
- training hardware and computational environment;
- exact software, framework, and package versions;

- original prediction file and complete multiclass confusion matrix;
- confirmed use or non-use of test-time augmentation during the reported evaluation; and
- confirmed use of the stored temperature value during the reported evaluation.

These missing records should not be reconstructed from default values in the available training scripts.

C. Training Configuration and Implementation Records

The project contains both a training script and a version-2 training workflow with configurable options for class-imbalance handling, augmentation, cross-validation, focal loss, test-time augmentation, and related training procedures. These configurations document available implementation capabilities but do not establish the historical training procedure of the evaluated checkpoint.

An important distinction is that the default train.py configuration specifies an EfficientNet-B3 backbone with 384×384 inputs, whereas the verified checkpoint is an EfficientNetV2-S model evaluated at 224×224 resolution. Consequently, these script defaults must be regarded as available experimental configurations rather than evidence of how the reported checkpoint was trained.

For future experiments, the complete training record should be retained, including dataset identity and version, class distribution, data partitions, preprocessing and augmentation operations, backbone configuration, loss function, optimizer, learning-rate schedule, batch size, number of epochs, random seed, hardware, software versions, training duration, checkpoint-selection criterion, and inference configuration.

D. Standardized Evaluation Protocol

All proposed configurations should be evaluated under identical primary conditions. The same held-out evaluation set should be maintained across E0–E7, and the evaluation partition should not be changed between experiments.

The classification evaluation should report:

- accuracy;
- balanced accuracy;
- macro-F1 and weighted-F1;
- per-class precision, recall, specificity, and F1-score;
- macro and weighted ROC-AUC;
- confusion matrix;
- malignant-class sensitivity using a clearly defined evaluation rule; and
- confidence intervals or other appropriate measures of uncertainty.

For segmentation experiments, Dice similarity coefficient and Jaccard index should be reported using the same reference masks and evaluation protocol for U-Net++ and TransUNet. For explainability experiments, Grad-CAM visualizations should be generated using the same model outputs and evaluated qualitatively and, where suitable

reference annotations are available, quantitatively. Computational measurements should additionally include parameter count, training time, inference time, and peak GPU memory consumption. These measurements are currently unavailable for the verified checkpoint and should therefore be collected during future controlled experiments.

E. Reproducibility Controls

Randomness should be controlled through a fixed seed policy. If multiple independent training runs are performed, results should be reported using the mean and variability across runs rather than selecting only the best-performing run. The prediction outputs used to calculate the reported metrics should be retained together with the corresponding ground-truth labels. Confusion matrices should be generated directly from these retained predictions. For segmentation experiments, both reference and predicted masks should be preserved so that Dice and Jaccard measurements can be independently reproduced. Each experimental configuration should also be associated with a clearly identified model checkpoint and a record of its training configuration. This prevents changes in preprocessing, data partitioning, or optimization settings from being incorrectly attributed to architectural modifications.

F. Reproducibility Boundary

The current study therefore establishes an inference-level reproducible EfficientNetV2-S baseline, but not a fully reproducible reconstruction of the historical training process. This distinction is important because the original training dataset, exact data partitions, and complete training configuration of the evaluated checkpoint are not sufficiently documented. The proposed U-Net++, TransUNet, EfficientNetV2–Vision Transformer fusion, and Grad-CAM components should consequently be considered planned experimental extensions until they are implemented and evaluated under the standardized protocol described above.

Once these experiments are completed, the complete experimental record should be retained so that the proposed framework can be compared directly with E0 without relying on undocumented assumptions. No improvement should be claimed solely from the proposed architecture; the contribution should be supported by controlled ablation results, uncertainty estimates, class-wise analysis, and, where possible, evaluation on an independent external dataset.

XIV. CONCLUSION AND FUTURE WORK

A. Conclusion

This study presented an EfficientNetV2-S-based approach for eight-class skin-lesion classification and established a verified baseline for further development of a computer-aided dermoscopic analysis system. The verified classifier was successfully loaded, executed, and evaluated on 120 images, achieving 61.67% accuracy, 55.47% balanced accuracy, 53.28% macro-F1, and 94.85% macro ROC-AUC.

The class-wise analysis showed that performance was not uniform across lesion categories, with stronger recognition of dermatofibroma and benign keratosis, while basal cell carcinoma and squamous cell carcinoma remained challenging.

The malignant-class analysis is particularly important. Using basal cell carcinoma, melanoma, and squamous cell carcinoma as the malignant categories, the verified model achieved an exact-class malignant sensitivity of 37.5%, with 25 of 40 malignant cases missed. This result demonstrates that the current classifier should be regarded as a research baseline rather than a clinically validated diagnostic system. Building on this baseline, the study proposed a segmentation-first hybrid framework combining lesion segmentation using U-Net++ or TransUNet, local feature extraction using EfficientNetV2, global contextual modelling using a Vision Transformer, and Grad-CAM-based visual interpretation. These components provide a technically motivated direction for improving lesion-focused representation and model interpretability; however, they remain proposed components and have not been assigned experimental performance in this study.

The work also demonstrates the integration of the verified classifier into a Flask-based web application supporting authentication, image analysis, prediction workflows, role-based review, and report generation. The application provides a practical research environment for subsequent model development while maintaining a clear distinction between automated prediction and clinical decision making. Overall, the study establishes a measurable baseline, identifies the major performance challenges, and provides a structured framework for subsequent experimentation. The reported results should therefore be interpreted as evidence of the current model's behaviour on the evaluated images, not as evidence of clinical effectiveness.

B. Future Work

Future development will focus on experimentally validating the proposed components and improving the reliability of malignant-lesion recognition. The main directions are: Implement and quantitatively evaluate U-Net++ and TransUNet for lesion segmentation.

Develop the EfficientNetV2–Vision Transformer feature-fusion architecture and compare it directly with the verified EfficientNetV2-S baseline.

Perform controlled ablation experiments to determine the individual contribution of segmentation, hybrid feature fusion, loss functions, and test-time augmentation.

Evaluate the proposed segmentation stage using Dice and Jaccard metrics against appropriate reference masks.

Integrate Grad-CAM and examine whether the highlighted regions correspond to clinically relevant lesion areas.

Perform a separate benign-versus-malignant evaluation in addition to the eight-class classification task.

Evaluate the system using independent datasets and patient- or lesion-level separation where appropriate to assess generalisation.

Investigate strategies specifically targeted at the low-recall classes, particularly basal cell carcinoma and squamous cell carcinoma.

Record computational characteristics including parameter count, inference time, training cost, and memory requirements to assess the practical trade-off introduced by the proposed hybrid architecture.

Conduct broader clinical and external validation before considering any clinical decision-support deployment.

C. Final Remarks

The principal outcome of this work is a transparent and experimentally verified baseline together with a structured research framework for its further improvement. The current results demonstrate both the potential and the limitations of deep-learning-based skin-lesion classification. In particular, the low malignant sensitivity indicates that improved classification reliability must remain a primary objective.

The proposed segmentation, hybrid CNN–Transformer, and explainability components provide a clear pathway for the next experimental stage. Their effectiveness should be established through controlled comparisons rather than assumed from architectural design alone. With rigorous validation, external testing, and appropriate clinical oversight, the framework can be further developed toward a more reliable and interpretable computer-aided skin-lesion analysis system.

REFERENCES

- [1] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Transactions on Medical Imaging*, 2020.
- [2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [3] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, DLMIA/MICCAI, 2018.
- [4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.
- [5] [Bibliographic details require verification from the original source.] "EFFNet: A skin cancer classification model based on feature fusion and random forests," *PLOS ONE*, 2023.
- [6] [Bibliographic details require verification from the original source.] "Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata," *MethodsX*, 2020.
- [7] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [8] [Bibliographic details require verification from the original source.] "Transformer-aided skin cancer classification using VGG19-based feature encoding," *Scientific Reports*, 2025.
- [9] R. R. Selvaraju, A. Cogswell, A. Das, A. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [10] 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.
- [11] 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, doi: 10.1038/nature21056.
- [12] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, and S. I. Karim, “Skin cancer segmentation and classification using Vision Transformer for automatic analysis in dermatoscopy-based noninvasive digital system,” *International Journal of Biomedical Imaging*, vol. 2024, Article ID 3022192, 2024, doi: 10.1155/2024/3022192.
- [13] C. Xin and Z. Liu, “An improved transformer network for skin cancer classification,” *Computers in Biology and Medicine*, vol. 149, Article 105939, 2022, doi: 10.1016/j.combiomed.2022.105939.
- [14] G. Yang, S. Luo, and P. Greer, “A novel Vision Transformer model for skin cancer classification,” *Neural Processing Letters*, vol. 55, pp. 9335–9351, 2023, doi: 10.1007/s11063-023-11204-5.
- [15] V. Anand, S. Gupta, D. Koundal, and K. Singh, “Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images,” *Expert Systems with Applications*, vol. 213, Part C, Article 119230, 2023, doi: 10.1016/j.eswa.2022.119230.
- [16] “InSiNet: A deep convolutional approach to skin cancer detection and segmentation,” *International Federation for Medical and Biological Engineering*, 2022.
- [17] “Modified U-NET architecture for segmentation of skin lesion,” *Computational and Mathematical Methods in Medicine*, 2022.
- [18] “Deep learning-based skin lesion multi-class classification with global average pooling improvement,” *Journal of Healthcare Engineering*, 2023.
- [19] “Skin cancer segmentation and classification using Vision Transformer for automatic analysis in dermatoscopy-based noninvasive digital system,” *International Journal of Biomedical Imaging*, 2024.
- [20] “Robust skin lesion segmentation in dermoscopic images using attention residual based UNet model,” *Procedia Computer Science*, vol. 283, pp. 4705–4714, 2026, doi: 10.1016/j.procs.2026.06.526.
- [21] A. M. A. Al-Masni, M. A. Al-Antari, M. T. Choi, S. M. Han, and T. S. Kim, “Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks,” *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 221–231, 2018.
- [22] M. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. E. Celebi, and G. Hamarneh, “A survey on deep learning for skin lesion segmentation,” *Medical Image Analysis*, vol. 88, Art. no. 102863, 2023, doi: 10.1016/j.media.2023.102863. ScienceDirect
- [23] H. A. Khan, Z. Ma, X. Chai, H. Qayyum, and others, “Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images: A review,” *Journal of Healthcare Engineering*, 2020. PubMed
- [24] A. J. M. A. T. et al., “Identifying the role of vision transformer for skin cancer—A scoping review,” *Frontiers in Artificial Intelligence*, 2023, doi: 10.3389/frai.2023.1202990. Frontiers
- [25] 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, doi: 10.1038/nature21056.
- [26] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. O. Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz, “A deep learning system for differential diagnosis of skin diseases,” *Nature Medicine*, vol. 26, pp. 900–908, 2020, doi: 10.1038/s41591-020-0842-3.
- [27] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, “Human–computer collaboration for skin cancer recognition,” *Nature Medicine*, vol. 26, pp. 1229–1234, 2020, doi: 10.1038/s41591-020-0942-0.
- [28] M. Goyal, M. S. Yap, N. N. L. H. et al., “InSiNet: A deep convolutional approach to skin cancer detection and segmentation,” *Medical & Biological Engineering & Computing*, 2022. The study evaluates its approach across ISIC/HAM10000-related datasets and compares it with architectures including DenseNet, ResNet, EfficientNet and conventional machine-learning models. PubMed
- [29] P. K. et al., “Classification of skin cancer from dermoscopic images using deep neural network architectures,” 2022. The study investigates EfficientNet transfer learning on ISIC-2019 and ISIC-2020 data and addresses class imbalance through augmentation and metadata. PubMed
- [30] M. Dakhli et al., “Explainable Deep Learning for Imaging-Based Skin Lesion Diagnosis: A Systematic Literature Review,” *Concurrency and Computation: Practice and Experience*, 2026, doi: 10.1002/cpe.70885.