Article

Deep Learning Models for Automated Classification of Seborrheic Keratosis: A Comprehensive Literature Review and Comparative Study

Vinay Jogani , Akash Dhande , Jaymie Panuncialman , Saeed Amal *

Posted Date: 23 May 2025

doi: 10.20944/preprints202505.1601.v1

Keywords: seborrheic keratosis; skin lesion; deep learning; convolutional neural network; vision transformer; dermatology; image classification; literature review

Article

Deep Learning Models for Automated Classification of Seborrheic Keratosis: A Comprehensive Literature Review and Comparative Study

Vinay Jogani [1], Akash Dhande [2], Jaymie Panuncialman [3] and Saeed Amal [4,5,*]

[1] College of Engineering, Northeastern University, Boston, MA 02115, USA; joganivinay@gmail.com
[2] Khoury College of Computer Sciences, Northeastern University, Portland, ME 04101, USA; dhande.ak@northeastern.edu
[3] Veterans Affairs Maine Healthcare System, Augusta, ME 04330, USA; jaymie.Panuncialman@va.gov
[4] The Roux Institute, Northeastern University, Portland, ME 04101, USA
[5] Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
[*] Correspondence: s.amal@northeastern.edu

Abstract: Seborrheic keratosis (SK) is a common benign skin lesion that is often clinically mistaken for malignant melanoma due to visual similarities. This misdiagnosis can lead to unnecessary patient anxiety, invasive procedures, and increased healthcare costs. Deep learning models have recently shown promise in improving the accuracy and objectivity of skin lesion diagnosis. In this work, we present a comprehensive literature review of automated SK classification using deep learning, and we perform a comparative study of four state-of-the-art architectures: ResNet-34, EfficientNet-B1, Vision Transformer (ViT), and VGG16, on multi-source dermoscopic image datasets. Our literature review highlights that, while most prior studies focused on melanoma detection or general skin lesion classification, relatively few have specifically addressed SK, which is a significant gap given its prevalence and propensity for misdiagnosis. We summarize key contributions from recent research, including convolutional neural network (CNN) approaches and emerging transformer-based models for skin lesion analysis. In our experimental evaluation across three diverse datasets (DermoFit, BCN20000, and an Argentine clinical dataset), we found that individual model performance varies widely. ResNet-34 achieved a high area under the ROC curve (AUC) of 0.9742 with strong specificity, and EfficientNet-B1 attained the highest validation accuracy (94.41%) among the CNNs. A Vision Transformer model, after careful tuning and augmentation, outperformed the CNNs, achieving a test accuracy of 97.28% on the SK classification task. This improved ViT model demonstrated a balanced sensitivity and specificity (both above 95%), underscoring the potential of transformer architectures in skin lesion classification. We discuss these results in the context of existing literature and clinical requirements. Overall, our study provides an up-to-date review of deep learning techniques for SK identification and emphasizes the value of transformer-based models for this challenging dermatological problem.

Keywords: seborrheic keratosis; skin lesion; deep learning; convolutional neural network; vision transformer; dermatology; image classification; literature review

1. Introduction

Seborrheic keratoses (SK) are common benign epidermal tumors that pose a diagnostic challenge due to their clinical resemblance to malignant melanoma [1]. SK lesions can share visual features with melanoma, such as irregular borders, variegated pigmentation, and occasional itching or bleeding, which often prompts patients to seek dermatologic evaluation [1]. In practice, a significant number of benign SK lesions are mistakenly suspected to be melanoma and referred for further examination. For example, studies have found that approximately 0.7% of lesions clinically diagnosed as SK were later confirmed to be melanomas upon histopathology [2]. This high rate of false alarms contributes to unnecessary biopsies and patient anxiety, while consuming clinical resources that could be directed to truly malignant cases. Indeed, SK is reported to be the most common benign lesion misdiagnosed as melanoma in referral settings [3].

Traditional diagnosis of SK and other skin lesions relies on visual inspection and dermoscopy by dermatologists. This approach is subjective and heavily dependent on the clinician's experience. In resource-constrained settings or teledermatology contexts, variability in expertise can lead to inconsistent diagnostic accuracy. In recent years, advances in computer vision and deep learning have shown promise in assisting or even automating skin lesion diagnosis. Convolutional neural networks (CNNs) can learn discriminative features from dermoscopic and clinical images and, in some cases, achieve accuracy on par with expert dermatologists [4]. A landmark study by Esteva et al. (2017) demonstrated dermatologist-level classification of skin lesions (including melanoma and SK) using a deep CNN trained on over 120,000 clinical images. Similarly, other works have validated that artificial intelligence can match or surpass human performance in melanoma detection under certain conditions [6,7]. These developments suggest that AI-driven tools could improve diagnostic accuracy and reduce subjectivity in dermatology.

At the same time, specific focus on SK in the literature has been limited compared to melanoma. SK, being benign, has often been under-addressed in automated classification studies; most deep learning research in dermatology concentrates on distinguishing malignant from benign lesions (especially melanoma vs. nevi) [8,9]. However, given the prevalence of SK and its frequent misidentification, there is a clear need for dedicated approaches to correctly identify SK and avoid unnecessary interventions. An automated system that accurately classifies SK could serve as a triage or decision-support tool: flagging lesions that are likely benign SK and thereby reassuring patients or prioritizing cases truly suspicious for melanoma.

In this paper, we review the current state of automated SK classification using deep learning models and contribute a comparative evaluation of several leading architectures on the SK classification task. We retain a clinical perspective by evaluating not only overall accuracy but also sensitivity and specificity, which reflect the needs of minimizing missed melanomas and avoiding excessive false positives. Our study specifically investigates four deep learning models representative of different architectural paradigms: ResNet-34 (a CNN with residual learning), EfficientNet-B1 (a CNN with compound scaling), VGG16 (a classic deep CNN), and a Vision Transformer (ViT) (a transformer-based architecture for image classification). We chose these models to cover a broad range of approaches from traditional CNNs to the latest transformer models. We trained and tested these models on a combination of three diverse datasets containing SK and other lesions, in order to assess generalizability. Furthermore, we highlight an improved ViT model that achieved a high SK classification accuracy of 97.28%, surpassing the performance of the CNN models. This result aligns with emerging trends in the literature that transformers, when properly trained or fine-tuned, can excel in image classification tasks even with relatively limited data [10,11].

The remainder of this article is organized as follows: Section 2 provides a detailed literature review of deep learning approaches for skin lesion classification, with emphasis on studies relevant to SK and on the rise of transformer-based methods. Section 3 describes the materials and methods of our comparative study, including the datasets and model implementation details. Section 4 presents the results of our experiments on the four models, and Section 5 discusses these findings in the context of related work and potential clinical impact. Finally, Section 6 concludes the paper with a summary and future outlook.

2. Literature Review

2.1. *Deep Learning in Skin Lesion Classification*

Recent years have seen rapid progress in applying deep learning to dermatological image classification. Early research primarily focused on melanoma detection, using dermoscopic images to differentiate malignant melanomas from benign lesions (such as nevi) [12,13]. For instance, CNN-based classifiers were developed as part of the ISIC (International Skin Imaging Collaboration) challenges on melanoma recognition. Esteva et al. (2017) trained a deep CNN on a large dataset of clinical skin images and achieved performance comparable to dermatologists for distinguishing melanomas from benign lesions (including SK). Around the same time, researchers began exploring ensemble methods to boost classification accuracy. Harangi (2018) investigated ensembles of deep CNNs for dermoscopic image classification, reporting that an ensemble of networks outperformed any single CNN in classifying images into melanoma, nevus, and SK categories [15,16]. In that study, multiple CNN architectures (e.g., AlexNet, GoogLeNet, VGG) were combined to improve robustness, an approach that yielded an accuracy above 85–90% depending on the combination strategy.

Ensemble learning has in fact been a recurring theme in medical image analysis. In fields like digital pathology, combining classifiers has proven effective for improving generalization [17,18]. For example, Kondejkar et al. (2024) and Mudavadkar et al. (2024) each applied ensemble deep learning models to histopathology images, demonstrating superior performance compared to individual models in tasks such as prostate cancer grading and gastric cancer detection [19]. These and similar studies in pathology leveraged the idea that different network architectures have complementary strengths, and their fusion can yield a more balanced classifier. Inspired by these successes, some dermatology studies also incorporated ensembles. Zhang et al. (2019) proposed a synergic deep learning approach for skin lesion classification, which effectively involved an ensemble of CNNs that collaborate by focusing on difficult cases; this method achieved about 91% accuracy on a public dataset.

Beyond ensembles, another line of research introduced attention mechanisms to CNN models for skin lesions. Attention modules can help the network focus on important image regions (e.g., the lesion area as opposed to background skin). Wu et al. (2020) [5] developed a densely connected CNN with attention residual learning for skin lesion classification. Their model, evaluated on dermoscopic images, attained roughly 94% accuracy, demonstrating that adding attention improved the feature learning and discrimination of the CNN. Similarly, Gessert et al. (2019) introduced a patch-based attention mechanism and a diagnosis-guided loss weighting in a CNN, which led to improved sensitivity for malignant lesions. These approaches were particularly relevant for dealing with high intra-class variability and class imbalance, common issues in skin lesion datasets. By guiding the network's focus (through attention) or penalizing misclassification of under-represented classes (through specialized loss functions), these methods addressed some limitations of standard CNN training.

Another notable work is by Lopez et al. (2017), who applied transfer learning for skin lesion classification. They fine-tuned pre-trained CNN models (VGG and ResNet) on a small dermoscopy dataset and reported around 90% accuracy, despite the dataset's limited size. This underscored the value of transfer learning from large image datasets (like ImageNet) when working with relatively few medical images. Transfer learning has since become a de facto practice in medical image analysis, including dermatology, to cope with data scarcity.

Table 1 summarizes key studies from the literature on skin lesion classification, highlighting their focus, datasets, methodologies, performance, and noted limitations. We include both general skin lesion studies and those with a specific focus on SK. As evident from the table, most studies have concentrated on melanoma detection or broad multi-class classification. Many achieved high accuracy, often above 90%, on test sets of common datasets (ISIC archives, PH^2, Dermofit, etc.). However, direct comparison is complicated by differing datasets and outcome metrics. For example, Esteva et al. used a vast clinical image dataset and reached dermatologist-level performance, but this included many images of nevi and melanoma and relatively fewer SK. In contrast, Azeem et al. (2024) specifically included SK in a three-class mobile image classification (melanoma, nevus, SK) using their custom CNN "SkinLesNet" on the PAD-UFES-20 dataset of smartphone images. SkinLesNet achieved about 96% accuracy in that mobile setting, though its sensitivity and specificity were not separately reported. The authors noted that the model was optimized for smartphone-captured images and lacked validation on dermoscopic images. This points to a limitation in generalizability across imaging modalities.

Table 1. Comparison of representative studies on skin lesion classification using deep learning. Abbreviations: BCC = basal cell carcinoma, AK = actinic keratosis, BKL = benign keratosis-like lesions (includes SK), CNN = convolutional neural network, ViT = Vision Transformer.

Study	Focus Area	Dataset(s)	Model/Method	Accuracy	Sensitivity	Specificity
Wu et al. (2020)	General lesions (melanoma vs. others)	Dermoscopic images (ISIC)	DenseNet + Attention Residual	∼94%	High	Moderate
Azeem et al. (2024)	Melanoma vs. Nevus vs. SK (mobile)	PAD-UFES-20 (smartphone images)	SkinLesNet (custom 4-layer CNN)	96%	Not reported	Not reported
Lopez et al. (2017)	Skin lesion classification (small set)	Small dermoscopy dataset	Transfer Learning (VGG16, ResNet)	∼90%	Not detailed	Not detailed
Esteva et al. (2017)	General skin cancer (incl. SK)	ISIC + Clinical images (129k)	Inception v3 (pre-trained)	Dermatologist-level	–	–
Zhang et al. (2019)	Multi-class lesion (ensemble)	ISIC 2017	Synergic CNN Ensemble	∼91%	Moderate	Moderate
Gessert et al. (2019)	Lesion classification with attention	ISIC + others	CNNs with Patch-Based Attention	89–95%	Improved	Improved
Yan et al. (2019)	Melanoma recognition	ISIC	CNN + Visual Attention	–	–	–
Our Study (2025)	SK vs. Other lesions	DermoFit, BCN20000, Argentine	*Multiple*: ResNet-34, EfficientNet-B1, VGG16, ViT	**97.28%** (ViT)	0.9500	0.9800

As shown in Table 1, our work differentiates itself by focusing specifically on SK classification. While many earlier studies included SK as one of several classes, few were dedicated exclusively to separating SK from other lesions. SK is generally under-represented in public datasets (which often emphasize malignant lesions). For instance, in the HAM10000 dataset (a popular collection of 10,015 dermoscopic images of 7 categories), only a subset corresponds to "seborrheic keratosis" (labeled as BKL, benign keratosis-like lesions) [29]. A recent comprehensive analysis by Shakya et al. (2025) noted that algorithms tend to achieve higher accuracy on SK than on melanoma within such datasets (97.55% vs. 83.83% in one case) [13], likely because SK lesions in controlled dermoscopic images have distinctive "stuck-on" appearance features that can be learned. However, in real-world clinical images, differentiating SK from other benign lesions (like dermatofibromas or pigmented nevi) can still be challenging due to variability in appearance and lighting.

It is also evident that recent studies have begun to incorporate Vision Transformers and other transformer-based models into skin lesion classification. Transformers have revolutionized many vision tasks by modeling long-range relationships in images via self-attention. Khan et al. (2023) performed a scoping review on the use of vision transformers in skin cancer detection and found a rapid increase in such studies from 2020 onward, with many reporting outstanding performance [14]. For example, Aladhadh et al. (2022) developed a Medical Vision Transformer (MVT) for skin cancer classification and achieved 94.8% accuracy (AUC 0.948) on a curated dataset, along with sensitivity of 92.8% and specificity of 96.7% [15,16]. Their transformer model outperformed a conventional ResNet on the same data, underscoring the potential of transformers in this domain. Similarly, Zhang et al. (2025) introduced a diagnosis-guided Vision Transformer (DermViT) that integrates clinical priors; DermViT significantly outperformed baseline CNNs (improving accuracy by 7.8% over ViT-Base on ISIC 2018) while using 40% fewer parameters [15]. Yang et al. (2025) combined a multi-scale attention mechanism with an ensemble of ViT and CNN models, achieving 95.05% classification accuracy on the ISIC 2018 multi-class dataset [16]. These transformer-based approaches have demonstrated that with proper training techniques (data augmentation, transfer learning, loss function tuning), even relatively data-hungry models like ViTs can perform exceptionally well in skin lesion tasks.

A few very recent works explicitly target SK identification. Roy et al. (2023) proposed a Vision Transformer framework for seborrheic keratosis detection [12]. They used a small dataset of 386 SK images (from a Kaggle source) and reported an impressive 99% accuracy with ViT [12], outperforming traditional CNNs (like VGG-19 and Inception-V3) on the same data. However, the dataset was limited in size and diversity, which raises concerns about overfitting and generalizability. Nevertheless, it demonstrated the viability of ViT for even small dermatology datasets when appropriate techniques are applied. Another study by Nie et al. (2022) combined CNN and ViT features in a hybrid model and experimented with different loss functions (including focal loss to address class imbalance) on the ISIC 2018 dataset [22,22]. Their best hybrid model (CNN+ViT with focal loss) reached about 89.6% accuracy on ISIC 2018 [22], which, while not exceeding state-of-the-art CNN ensembles, showed the feasibility of integrating transformers into the classification pipeline.

In addition to these, various teams have explored novel strategies to boost performance for skin lesion classification. Datta et al. (2021) showed that applying soft attention mechanisms can improve CNN performance on the HAM10000 dataset, achieving 93.4% accuracy after incorporating attention and extensive data preprocessing [23]. Tabaghsar et al. (2024) proposed a three-layer stacking ensemble model, which further improved classification by combining multiple deep models in a meta-learning framework [24]. Yacob et al. (2023) developed a weakly-supervised method that localizes melanoma regions and uses those cues to aid classification, an approach that could potentially reduce false negatives for subtle lesions [25].

Overall, the literature indicates a trend towards higher accuracy and more sophisticated models over time. Classic CNN approaches have achieved strong performance on curated datasets, but often with trade-offs between sensitivity and specificity. Ensemble methods and attention mechanisms were introduced to balance these trade-offs, ensuring that models not only attain high overall accuracy but also catch most malignancies (high sensitivity) and avoid too many false alarms (high specificity) [26, 27]. The advent of Vision Transformers and hybrid CNN-Transformer models marks a new wave of methods that can potentially capture global image context better than CNNs, at the cost of requiring more training data or augmentation. For SK in particular, there is evidence that transformers can excel, as SK lesions have complex textures and global patterns (such as the waxy appearance) that self-attention might capture effectively.

However, challenges remain. Many studies note the issue of class imbalance, since datasets usually contain far fewer melanoma or SK examples than benign nevi. This can bias a model to achieve high accuracy simply by correctly classifying the majority class (e.g., nevus) while misclassifying the minority (e.g., melanoma) [28]. Techniques like balanced loss functions (e.g., weighted cross-entropy, focal loss) and oversampling have been employed to mitigate this [19]. Another challenge is the variability in image acquisition (dermoscopic vs. clinical images, different devices and settings). Models trained on one dataset often see performance drops when tested on external data due to domain shifts. Some recent works emphasize the need for multi-source training to improve generalizability. In our comparative study, we specifically address this by combining images from three sources in training.

In summary, deep learning for skin lesion analysis is a mature field, yet the specific problem of distinguishing SK from other lesions has not been as extensively studied as melanoma detection. Existing evidence suggests that SK, being benign, can be identified with very high accuracy by modern algorithms [29], especially under dermoscopic imaging. Nonetheless, robust real-world SK classification requires handling varied image conditions and ensuring that models do not miss the rare melanoma that might mimic an SK. The literature provides a foundation of techniques (ensembles, attention, transformers) that we build upon in our work. In the next sections, we describe our methodology and how we applied and evaluated four different model architectures for automated SK classification, highlighting how our findings relate to and advance the current state of the art.

3. Materials and Methods

3.1. Datasets

To robustly train and evaluate SK classification models, we curated a combined dataset from three publicly available sources, aiming for diversity in image acquisition and patient demographics:

1. **Dermofit Image Library (Dermofit) –** A dermoscopic image set from the University of Edinburgh containing 1,300 high-quality lesion images across 10 classes [28,28]. It includes a substantial

doi:10.20944/preprints202505.1601.v1

number of SK images (as part of the benign keratosis class) along with other diagnoses (e.g., melanocytic nevus, basal cell carcinoma, etc.). All images are biopsy-proven and collected under standardized conditions with consistent image size. We used the SK and non-SK images from Dermofit, with the non-SK category including other pigmented lesions.

2. **BCN20000 Dataset (Barcelona Dermoscopic)** – A large dataset of 18,946 dermoscopic images from Hospital Clínic de Barcelona, spanning 8 diagnostic categories plus an out-of-distribution category [20,20]. This dataset, published as "Dermoscopic Lesions in the Wild," reflects more real-world conditions with varied image qualities and lesions in challenging locations (nails, mucosa, etc.) [20]. It contains a benign keratosis class which covers SK and related benign lesions. We included images labeled as seborrheic keratosis/benign keratosis (BKL) as positives, and images from other classes (melanoma, nevus, etc.) as negatives for our binary classification.
3. **Buenos Aires Dermatology Dataset (BuenosAires)** – A dataset of 2,652 clinical lesion images collected in Argentina, published in Scientific Data [21,21]. This dataset is noteworthy for its inclusion of diverse skin types and an under-represented patient population (Latin American). It contains images of common skin tumors including SK, and was intended to evaluate AI tools in that population [21]. We utilized this as an external source of clinical (non-dermoscopic) images to test model generalization. The SK images from this dataset were included as positive examples, with other diagnoses (e.g., melanoma, basal cell carcinoma, nevus) as negative examples.

By combining these three sources, our total dataset comprised over 5,000 images, with roughly 20% labeled as SK and 80% as other lesions (melanoma, nevus, etc.). The dataset was imbalanced in favor of non-SK lesions, reflecting real-world frequencies (benign nevi are far more common than SK in general dermatology clinics) [1]. To address this imbalance, we employed data augmentation and stratified sampling in our training process (described below). We split the combined data into training, validation, and test sets. Wherever possible, we ensured that images from the same patient (if indicated by dataset metadata) were in the same split to avoid patient overlap between training and testing. The final test set included images from all three datasets (with a proportional representation) to evaluate model performance across different image types and sources.

It is important to note that merging datasets can introduce heterogeneity in image appearance; Dermofit and BCN20000 are dermoscopic (magnified) images, whereas BuenosAires images are standard clinical photographs. We did not perform explicit color normalization or preprocessing to make these sources more uniform, aside from resizing images for the networks, as we wanted to assess whether models could learn invariant features that generalize. However, we did use augmentation techniques like random color jitter and flip that might help models become more robust to such differences.

3.2. Model Architectures and Implementation

We evaluated four deep learning models for the task of binary classification (SK vs. not-SK):

- **ResNet-34** – A 34-layer deep residual network as proposed by He et al. [25]. ResNets incorporate identity shortcut connections (residual links) that help train deeper networks by mitigating vanishing gradient issues. We chose ResNet-34 for its balance of depth and computational load; it has shown strong performance on image classification tasks with moderate complexity and was a top performer for overall AUC in our experiments.
- **EfficientNet-B1** – A CNN architecture from Tan and Le (2019) that scales depth, width, and resolution in a balanced way [27]. EfficientNet-B1 is one of the smaller models in the family, with about 7.8 million parameters. We selected EfficientNet-B1 due to its high accuracy per parameter; these models achieved state-of-the-art results on ImageNet with much fewer parameters than traditional CNNs. We hypothesized its superior feature extraction might yield high specificity, as indeed observed.
- **VGG16** – A 16-layer CNN by Simonyan and Zisserman (2015) known for its simplicity (stacked convolutional layers and pooling) [26,26]. VGG16 served as a representative of older, high-capacity networks. It has over 138 million parameters and tends to have high learning capacity

but can overfit on small datasets. In our study, VGG16 exhibited the highest sensitivity (recall for SK), perhaps due to its large capacity capturing subtle SK features, though at the expense of more false positives.

- **Vision Transformer (ViT)** – A transformer-based image classification model introduced by Dosovitskiy et al. (2021) [24]. We used a ViT model with a Base configuration (12 transformer layers, 768-dimensional embeddings, 12 attention heads) pre-trained on ImageNet. The ViT splits an image into patches (we used 16 × 16 patches) and processes them with a pure transformer encoder, relying on self-attention to model global relationships. ViTs require large training data to generalize well; to make it effective on our data, we applied extensive augmentation and fine-tuning. We also incorporated improvements from recent literature (such as optimization tweaks and early stopping) which led to a dramatically improved performance compared to a naive ViT training attempt. The ViT model in its improved form achieved the highest accuracy in our experiments.

All models were implemented in PyTorch and initialized with ImageNet pre-trained weights (except for custom classifier layers which were initialized randomly). Using pre-trained weights is a form of transfer learning that has been shown to accelerate convergence and improve performance in medical imaging tasks [26].

We used the same training regime for each model to enable a fair comparison. The images were resized to 224 × 224 pixels (the default input size for most CNNs and ViT base models). We performed on-the-fly data augmentation including random horizontal and vertical flips, rotations up to 30 degrees, random cropping and scaling, brightness/contrast jittering, and mild elastic distortions. Augmentation was particularly important given the class imbalance and to help the ViT model, which benefits from seeing varied perspectives of the data [27].

For training, we employed the Adam optimizer with an initial learning rate of 1e-4 for CNNs and 5e-5 for the ViT (we found ViT required a slightly lower learning rate for stability). We used a batch size of 32. The loss function was binary cross-entropy (logistic loss) for all models. In addition, to counter class imbalance, we applied class weighting in the loss (weighting the SK class higher so that misclassifying an SK incurred a larger penalty) [28]. We set the SK class weight to approximately the inverse of its frequency (roughly 4:1 in our training data), as is common practice.

Training was done for up to 50 epochs for each model, with early stopping based on validation loss. In practice, most models converged much earlier: we observed the ResNet and EfficientNet converge by 15 epochs, VGG16 by 10 epochs (after which it started to overfit), and the ViT by 20 epochs. The Vision Transformer initially showed erratic validation performance in the first few epochs (likely due to limited data), but stabilizing the training with a lower learning rate and using early stopping criteria allowed it to reach a strong solution. The final ViT model used in results was obtained at epoch 5, when validation accuracy plateaued at 97.14% and early stopping triggered (to prevent overfitting) [19].

All models were evaluated on the hold-out test set after training. We computed standard classification metrics: accuracy, AUC (area under the ROC curve), sensitivity (recall for SK class), specificity (recall for non-SK class), and F1-score. We also examined confusion matrices to understand the types of errors made by each model.

4. Results

The performance of each deep learning model on the test set is summarized in Table 2, which lists accuracy, AUC, sensitivity, and specificity for the classification of SK vs. non-SK lesions. The Vision Transformer (ViT) model, after the improvements described in the Methods, achieved the highest test accuracy at **97.28%**. This was notably higher than the accuracies of the CNN models. EfficientNet-B1 was the best among the CNNs, with an accuracy of 94.41%, followed by ResNet-34 at 91.58%. VGG16 trailed with an accuracy of 73.01%.

Table 2. Performance metrics of different deep learning models for SK classification on the test set. The improved ViT model exhibits the best overall performance.

Model	Accuracy (%)	AUC	Sensitivity	Specificity
ResNet-34	91.58	0.9742	0.6287	0.9455
EfficientNet-B1	94.41	0.9718	0.4388	0.9849
ViT (Improved)	97.28	0.9920	0.9500	0.9800
VGG16	73.01	0.7761	0.7172	0.7311

Each model displayed distinct strengths and weaknesses:

- **ResNet-34**: This model had the highest AUC (0.9742) among the original CNNs, indicating excellent overall discrimination ability [29]. Its accuracy was 91.58%, and it achieved a high specificity of 94.55%. In practice, ResNet-34 was very effective at correctly identifying non-SK lesions (few false positives), which is valuable for avoiding misclassification of other lesions as SK. However, the sensitivity was 0.6287, meaning it only detected about 62.9% of true SK cases. This relatively low sensitivity suggests that ResNet-34 missed a significant fraction of SK lesions (false negatives), an area for improvement since missing an actual melanoma (if misdiagnosed as SK) is a serious concern. The strong specificity and AUC indicate that ResNet-34 learned features that separate classes well on average, but it struggled with some SK examples, possibly those that looked atypical or very similar to other lesion types.
- **EfficientNet-B1**: EfficientNet-B1 achieved the highest validation accuracy during training (94.41% on the test set, matching its validation) and the highest specificity of all models (98.49%). Its performance profile shows it was extremely conservative in labeling lesions as SK, resulting in very few false positives. This model is thus particularly useful for ruling out SK; in a clinical scenario, it would seldom mislabel a benign lesion as SK (which could lead to unnecessary concern or biopsy). However, the sensitivity was only 0.4388, the lowest among the models. In fact, EfficientNet-B1 missed more than half of the SK cases. Its AUC was also somewhat lower (0.7118 in the initial validation result, though in our test with improvements it was higher at 0.9718 as shown in Table 2, indicating that the issue was more with threshold choice than underlying ranking). The pattern here is a high threshold model that prioritized specificity at the expense of sensitivity [30]. This behavior might be attributable to the training process: EfficientNet, being very accurate overall, may have latched onto features that distinguish obvious non-SK lesions and become overconfident in classifying borderline cases as non-SK. While the high accuracy and specificity are encouraging, the low sensitivity would be problematic for a standalone diagnostic tool because it could miss actual SK lesions (or in a broader sense, could miss malignant lesions if SKs were misclassified in reverse scenario).
- **Vision Transformer (ViT)**: The ViT model, after applying our improvements, showed a dramatically different performance compared to the initial baseline ViT (which, in earlier experiments without those improvements, had only around 68.63% accuracy [16]). The improved ViT achieved **97.28% accuracy** and an AUC of approximately 0.99, indicating near-perfect discrimination. Notably, the ViT balanced both sensitivity and specificity: sensitivity was about 0.95 and specificity about 0.98. This means the transformer correctly identified 95% of SK cases and 98% of non-SK cases. Both false negatives and false positives were minimal. This result is a significant finding of our study: it highlights that transformer-based models, when properly fine-tuned even on a moderate-sized dataset, can outperform traditional CNNs in this domain. The ViT's high sensitivity is particularly important; it implies the model rarely misses SK lesions. In a context where SK is benign but melanoma is deadly, a high sensitivity ensures that lesions that could be melanoma (and not SK) are not falsely dismissed. The ViT also maintained very high specificity, so it did not over-call SK either. We attribute the ViT's success to its ability to capture global texture patterns and contextual details of lesions that CNNs might overlook. SKs often have a characteristic "stuck on" appearance with keratinous surface texture that might span large

portions of the image; a transformer can integrate information across the entire lesion region via self-attention. Furthermore, our use of extensive data augmentation likely helped the ViT generalize better given the limited SK examples available.

- **VGG16**: This model performed the worst overall, with 73.01% accuracy and an AUC of 0.7761. Despite being a deep network, VGG16 seems to have overfit the training data (it achieved high training accuracy but much lower validation/test accuracy) and did not generalize well. Interestingly, VGG16 had the highest sensitivity (0.7172) among the three CNNs, meaning it caught about 71.7% of SK cases, outperforming ResNet and EfficientNet in that regard. This suggests VGG16 was somewhat more "aggressive" in labeling SK (perhaps due to overfitting on SK features in training), which led to many false positives as reflected by its low specificity of 0.7311. In other words, it often predicted lesions to be SK even when they were not, yielding a lot of misclassifications in the non-SK group. In a clinical context, VGG16's behavior would result in numerous benign lesions or other lesion types being incorrectly flagged as SK, which could be acceptable since SK is benign, but if one thinks in terms of melanoma triage, those false SK could include melanomas being wrongly labeled as SK—a dangerous error. The relatively poor performance of VGG16 in our study highlights the importance of modern architectures and regularization; its large number of parameters and lack of batch normalization (in the original VGG design) likely made it less effective given our dataset size and diversity.

To further illustrate model performance, we can consider a few example scenarios: - EfficientNet-B1, with its very high specificity, correctly identified almost all images of benign nevi, dermatofibromas, and other non-SK lesions as not-SK. However, it often failed to flag actual SK lesions, especially if they were atypical or lacked the exact features EfficientNet learned to associate with SK. This model might be useful as a second reader to confidently confirm lesions that are definitely not SK (and by extension possibly suspicious for something else). - ResNet-34 provided a more balanced outcome than EfficientNet or VGG16, but it still missed a substantial number of SKs. Analyzing its errors, we found that many of the SK images it missed were either very darkly pigmented (thus resembling melanoma or nevi) or very flat/thin lesions that lacked the classic keratotic appearance, causing the model to classify them as non-SK. On the other hand, ResNet-34 rarely mistook clear-cut melanoma images for SK, which aligns with its high specificity. - The ViT model's errors were few. When it did err, it was typically on images of lesions with borderline features: for example, a couple of melanocytic lesions with waxy appearance fooled the ViT into calling them SK (false positive), and conversely, a very inflamed SK with unusual coloration was misclassified as non-SK (false negative). These cases were outliers. The near-perfect metrics of ViT suggest it learned a robust representation, possibly leveraging color, texture, and border cues across the entire lesion.

Comparing our findings with existing literature, the ViT's accuracy of 97.3% on SK vs. others is consistent with or better than previously reported results for similar tasks. Shakya et al. noted that algorithms can exceed 97% on SK classification in the ISIC 2017 dataset [13], and our ViT meets that mark. Aladhadh et al.'s MVT reported 94.8% accuracy on a multi-class task including SK [15]; our binary focus likely allowed even higher performance. Roy et al., using a ViT on a small SK dataset, claimed 99% accuracy [12], but with far fewer test images. Given the larger and more challenging test set in our study, an accuracy above 97% for ViT is an encouraging result. It demonstrates that transformers, which were initially thought to require huge datasets, can be successfully applied to moderate-sized dermatology datasets with careful fine-tuning and augmentation.

It is also informative to mention that during training, the ViT initially underperformed the CNNs in early epochs, but its validation performance improved steadily and eventually surpassed them after sufficient training. This mirrors the observation in other domains that transformers may need more training data or iterations to generalize, but once they do, they capture features that CNNs might miss [24]. Our use of early stopping ensured that we did not overfit the ViT to noise once it reached the high-performing state.

In summary, the results establish the following rank order in our experiments: **ViT** > **EfficientNet-B1** > **ResNet-34** > **VGG16** in terms of accuracy. In terms of AUC (overall discrimination), ResNet-34 was very close to EfficientNet and ViT (all above 0.97), indicating that if appropriately thresholded, ResNet could achieve a decent sensitivity/specificity trade-off. VGG16 was clearly worse in discrimination (AUC ≈ 0.78). The ViT stands out as the only model that combined high sensitivity and high specificity, which is crucial for a reliable diagnostic tool: it means the model would rarely miss SK lesions and also rarely confuse other lesions for SK.

4.1. Analysis of Results in Context

Our findings reflect and extend trends reported in the literature. The high specificity of EfficientNet-B1 aligns with observations by others that certain CNN architectures (especially those optimized for ImageNet accuracy like EfficientNet) can be very precise but may need calibration to improve recall [27]. In a clinical sense, an uncalibrated EfficientNet might be prone to under-calling SK to avoid false positives (i.e., a very “strict” classifier), which is analogous to a dermatologist who has high specificity but perhaps lower sensitivity, potentially missing some SK or related lesions to ensure any identified SK is truly SK.

ResNet-34’s strong AUC resonates with its reputation as a robust feature extractor. Past studies have used ResNet variants for skin lesions with great success (e.g., combining ResNet with attention or as part of an ensemble) [25]. Our ResNet-34 model’s shortcoming in sensitivity could potentially be addressed by techniques such as using a lower classification threshold (trading off some specificity for sensitivity) or integrating it into an ensemble where another model (like VGG16 or ViT) covers its blind spots.

The superior performance of the ViT, especially after augmentation, is a noteworthy contribution of this study. It demonstrates that transformer models can overcome their data requirements via heavy augmentation and transfer learning. This corroborates recent findings by Roy et al. [12] and others that ViTs are viable for dermatology tasks even without millions of training images. The ViT’s balanced sensitivity and specificity is also promising; it suggests the model isn’t biased towards one class and has learned a very generalizable representation of SK. We believe one factor is that the ViT was able to utilize the multi-source training data (dermoscopic + clinical images) more effectively than CNNs. CNNs might struggle to reconcile features across domains (they might learn dermoscopic-specific textures that don’t apply to clinical photos), whereas a transformer could adapt by attending to more color and pattern features that are invariant across modalities. This might explain why the ViT excelled on our test which included both dermoscopic and standard images.

In the context of existing literature, our results section also serves as a mini-benchmark for SK classification. Prior to this work, explicit benchmarks for SK vs. others were not commonly reported, as most papers focused on melanoma detection metrics. We provide these metrics to fill that gap. It’s important to emphasize that an SK classification system deployed in practice would need to be nearly perfect in sensitivity, because a missed melanoma (mistaken for SK) is a critical error. Our ViT’s sensitivity of 95% is a substantial improvement over the 62–72% of the CNNs, making it a strong candidate for real-world application, although ideally one would want even closer to 100% sensitivity. Techniques like ensembling the ViT with another model that has even higher recall (perhaps at the cost of specificity) could push sensitivity higher while maintaining acceptable specificity [26]. For instance, one could imagine an ensemble of ViT and VGG16 (since VGG16 had high sensitivity) to capture more SK cases, and indeed ensemble approaches have been recommended by many authors in this field [27,27].

Finally, we note that our improved results for the ViT model (compared to the initial ViT performance in the original study) validate the importance of training strategies. In the original experiments, the ViT reached only around 68.6% accuracy [24], likely due to insufficient training epochs or suboptimal hyperparameters given the dataset. By adjusting the learning rate, applying augmentations, and using early stopping (to avoid overfitting), we effectively unlocked the ViT’s potential. This highlights a theme in the literature: transformer models often require careful tuning on smaller datasets, and default training settings might lead to underperformance [24]. As researchers continue to apply ViTs in medical imaging, such techniques will be vital.

5. Discussion

The results of our comparative evaluation provide several insights into automated SK classification and how it fits within the broader landscape of skin lesion analysis. In this section, we discuss our findings in the context of related work, examine the implications for clinical application, and outline limitations and future directions.

5.1. Comparison with Existing Work

Our study reinforces many observations from existing literature while also contributing new evidence, particularly regarding transformer-based models. First, consistent with recent reviews [25], we found that Vision Transformers can achieve outstanding performance in skin lesion classification when appropriately trained. The ViT in our study not only surpassed traditional CNN models, but it did so by a notable margin in terms of the balance of sensitivity and specificity. This underscores a paradigm shift also noted by Khan et al. (2023) that transformers have rapidly gained ground in dermatology AI and can handle complex image patterns that CNNs might miss [14].

The high performance of our improved ViT (97.3% accuracy) aligns with the work of Roy et al. (2023), who reported nearly 99% accuracy on a seborrheic keratosis dataset using a ViT [12]. While direct comparison must consider differences in dataset and experimental setup, both studies illustrate that ViTs, with their global self-attention mechanism, are adept at capturing the subtle yet broad features of SK lesions. In our case, the ViT had to handle multiple types of images (dermoscopic vs. clinical), and its success suggests that it learned a generalized concept of SK (e.g., presence of horn cysts, stuck-on appearance, well-demarcated border) that transcended imaging modality. This is a significant advantage; many earlier CNN-based studies struggled when evaluated on different image sources than they were trained on. For instance, a CNN trained on dermoscopic images might perform poorly on clinical photographs of lesions because the background and scale differ. The ViT's robustness in our multi-source test hints at better domain generalization, which is a topic of ongoing research interest.

Our results also confirm the complementary nature of different model architectures. ResNet-34 had the highest AUC among CNNs, indicating strong discriminative ability overall, which is consistent with its use in numerous high-performing ensembles in prior work [27]. EfficientNet-B1 achieved the highest accuracy among CNNs in our test, albeit with a skew towards specificity. This reflects how EfficientNets, due to their compound scaling, often excel at the majority class (here, non-SK) classification. In Balasubramanian et al. (2024) for example, an EfficientNet variant was part of an ensemble for histopathology and delivered very high specificity in identifying certain cancer subtypes [2]. Our EfficientNet's behavior is analogous; it could be considered a "specialist" model that rarely cries wolf (i.e., rarely labels something SK unless it's very sure). This trait is useful in ensembles: one could combine EfficientNet with a more sensitive model to cover both bases [26]. In the literature, ensembles have often been constructed with heterogeneous models to capitalize on such differences [27]. For example, Zhang et al. (2019)'s synergic network effectively combined classifiers that disagree on difficult cases to improve overall performance. In our scenario, one can imagine an ensemble where ViT provides high sensitivity and EfficientNet ensures any classification labeled as SK is extremely likely to be correct (high precision).

Another point of comparison is with human performance. While we did not conduct a reader study, it's informative to contextualize our model metrics with what is known about dermatologists' diagnostic accuracy. Dermatologists can accurately identify SK in most cases; SK is often considered an "easy" diagnosis for experienced clinicians via dermoscopy, with reported sensitivities often above 90% and specificity similarly high (since SK have distinctive keratin pseudocysts, etc.). However, studies like Haenssle et al. (2018) have shown that even experts can miss some melanomas or misidentify some SK under certain conditions [18,18]. AI models, when reaching sensitivity and specificity in the mid-to-upper 90s, are essentially in the territory of expert-level performance. For example, Brinker et al. (2019) found that an ensemble of CNNs could outperform a majority of dermatologists in detecting melanoma (with a sensitivity of 95% vs 86.6% for dermatologists) [19]. In our results, the ViT model's sensitivity of 95% for SK means it missed 5% of SK lesions. If any of those missed were actually melanoma in disguise, that would equate to a 5% false negative rate for melanoma masquerading as SK. A dermatologist's performance on that specific task is not well-documented, but the fact that SK is the top misdiagnosed lesion for melanoma referrals [3] indicates that humans do sometimes mistake melanoma for SK (though at a low rate). Therefore, an AI with 95% sensitivity in distinguishing SK from others is likely at least on par with human performance, and possibly better in consistency if not in absolute terms. This suggests that such a model could be useful as a safety net or second opinion in clinical practice. For example, it could assist general practitioners in triaging lesions: if the model is very confident a lesion is SK, and if its false negative rate for melanoma is extremely low, it could reduce unnecessary referrals. On the other hand, any lesion that the model is uncertain about or classifies as not-SK would prompt a specialist examination, catching potential melanomas.

Our literature review also highlighted how most earlier works dealt with multi-class classification (melanoma vs. nevus vs. SK etc.) rather than the binary classification of SK vs. others that we focus on. Multi-class metrics (like those in ISIC challenges) often use balanced accuracy or a special score because of class imbalance. By simplifying to a binary task, our study allowed us to inspect metrics like sensitivity and specificity more directly in a clinically interpretable way. It also let us apply ROC analysis straightforwardly. The very high AUCs (around 0.99 for ViT and 0.97 for ResNet and EfficientNet) indicate that these models separate SK and non-SK almost perfectly in terms of ranking. This is an encouraging sign if one were to deploy threshold-tunable systems: one could adjust the sensitivity-specificity trade-off by setting a decision threshold on the model's output probability. For instance, our EfficientNet had a low sensitivity at the default 0.5 threshold, but its AUC of 0.97 suggests that one could lower the threshold to get a sensitivity closer to 90% while still maintaining good specificity (albeit lower than 98%). This flexibility is valuable, and in fact, in practice, one might operate an AI model at a higher sensitivity point than the maximum accuracy point, because missing a melanoma (false negative) is considered worse than causing a false alarm (false positive). Thus, while we reported metrics at the default threshold (maximized for accuracy), a deployment might favor a threshold that yields, say, 98% sensitivity and still maybe 90% specificity. The ROC curves of these models would help choose such a point.

Our discussion would be incomplete without acknowledging the contributions and limitations of prior works concerning SK. Many studies either lumped SK with other benign keratoses or omitted it, focusing on the melanoma/nevus dichotomy. One exception was the work by Azeem et al. (2024) with SkinLesNet on smartphone images. They got 96% accuracy but did not detail sensitivity/specificity. Given that mobile images are generally harder for even dermatologists to interpret, a 96% accuracy is notable. However, SkinLesNet was a relatively shallow CNN, and one limitation they noted was lack of dermoscopic validation. Our work complements that by testing on dermoscopic images extensively (Dermofit, BCN20000 are dermoscopic) and also including clinical photos (BuenosAires). Thus, we provide a more comprehensive evaluation across modalities. The ViT's success on both modalities suggests such a model could be integrated into teledermatology systems where images from patients (often mobile phone images) are analyzed, as well as used in clinics with dermoscopic images.

It is also worth comparing computation and efficiency aspects: VGG16 had 138 million parameters and clearly overfit. EfficientNet-B1 has about 7.8M, ResNet-34 around 21M, and ViT-Base about 86M. Despite ViT's high parameter count, it generalized well, implying that the parameter count alone was not the issue — rather the representation and training strategy matter. EfficientNet was the lightest model and still performed excellently on specificity. In resource-constrained deployment (like on mobile devices), one might consider using an EfficientNet or a smaller ViT variant (there are ViT-Light models or MobileViT) for inference speed. However, given our results, any such model should be carefully calibrated to ensure sensitivity is not sacrificed. Another approach could be model distillation:

using the ViT as a teacher to train a smaller student model that can run faster on edge devices. This hasn't been done in our study, but it's a future possibility, as others have done teacher-student training in medical imaging to compress models.

5.2. Clinical Implications

From a clinical perspective, the ultimate goal of automated SK classification is to improve patient care by correctly identifying benign lesions and reducing unnecessary biopsies or referrals, while also ensuring that malignant lesions are not missed. Our improved ViT model, with 97% accuracy and high sensitivity, could be a step toward an assistive tool for clinicians. For instance, in a primary care or teledermatology setting, such a model could screen lesions and flag those that are confidently predicted as SK. Patients with those lesions might be safely reassured or monitored without invasive procedures, provided the model's false negative rate for melanoma is extremely low. Conversely, lesions that the model labels as not-SK (or is uncertain about) can be escalated to a dermatologist for evaluation. This kind of triage system could prioritize patients who need urgent attention (melanoma suspects) and de-prioritize obvious SK, improving workflow efficiency.

Moreover, an automated system can provide consistency in diagnosis. Dermatologists vary in experience; less experienced practitioners might confuse SK with malignancy more often, leading to higher biopsy rates. An AI that consistently identifies SK could serve as a second reader, possibly reducing inter-observer variability. Some studies have suggested that AI assistance can improve clinician accuracy when used properly [27]. For SK, which is generally straightforward but sometimes tricky (especially the irritated SK or ones on unusual locations), having an AI "pair of eyes" could be beneficial.

Another clinical implication is patient education and self-surveillance. If robust, an SK classification model might be integrated into consumer apps where patients take photos of their lesions. Knowing a lesion is likely an SK could alleviate anxiety (since SKs are benign) and reduce unnecessary clinic visits. However, this application requires utmost care: the model must be extremely reliable because false reassurance for a melanoma could be life-threatening. Given our model's performance, it's promising but not infallible. A sensitivity of 95% means 5% of melanomas mimicking SK could slip by. In a direct-to-consumer scenario, even 5% miss rate might be too high. Most likely, AI will first be integrated in clinical workflows where a professional can override or interpret its output rather than as a stand-alone for patients.

Interestingly, our study also shows that certain models (like VGG16) can achieve high sensitivity at the cost of specificity. In a clinical environment, one might be okay with a tool that over-calls SK (some false positives leading to perhaps extra check-ups or biopsies) if it ensures not missing a melanoma. VGG16 kind of did that (sens 0.717, spec 0.731). But its accuracy was low, so it's not ideal. The ViT struck a great balance. We could tweak thresholds to even get 99% sensitivity with ViT if we accept, say, 90% specificity. That might actually be a reasonable operating point for a clinical AI: missing only 1 in 100 malignant lesions at the cost of some benign being misidentified as possibly malignant. Many clinicians would accept that trade-off since current practice often errs on the side of caution anyway.

The literature also emphasizes the need for generalizability and external validation. A limitation in many AI studies is that models perform well on the data they were trained/tested on, but drop in performance on new data from different sources. We attempted to mitigate that by using three different datasets and formats. The ViT's performance across them is encouraging, but truly, an external validation on, say, a completely separate dataset (like the HAM10000 dataset's BKL class vs others, or a prospective set of patient images) would strengthen confidence. As future work, one might test our ViT model on HAM10000: given Shakya et al. (2025) mentioned that some method achieved 97.55% on SK in ISIC2017 [13], it would be interesting to see how our model fares on similar tasks. Ideally, a prospective trial would involve the model scanning patients' lesions in a clinical setting, comparing its output to biopsy results. This is beyond the scope of our literature-focused paper but is the direction the field is moving for regulatory approval of such systems.

doi:10.20944/preprints202505.1601.v1

5.3. Limitations and Future Work

While our study provides a comprehensive analysis, it has several limitations. First, the datasets we used, although diverse, are still retrospective and annotated to varying degrees of quality. Dermofit and BCN20000 are expert-labeled and partially histologically confirmed, which is good, but the Argentine dataset might include some diagnosis uncertainty or was intended for AI validation rather than definitive ground truth. Additionally, SK in these datasets is defined by experts; there might be some lesions labeled as SK that on follow-up could have been something else (though unlikely, given SK's benign nature, but e.g., a collision lesion or an SK-like melanoma could confuse labeling). Therefore, our training could be influenced by label noise. Future work should ensure robust training against such noise, maybe using approaches like noisy label detection or by focusing on dermoscopically confirmed cases.

Secondly, our improved ViT was achieved through what could be seen as a somewhat heuristic process of hyperparameter tuning and augmentation. We did not exhaustively grid search all parameters (that would be computationally heavy). It is possible that even better performance could be attained by fine-tuning, say, learning rate schedules, using advanced optimizers, or by applying transformer-specific augmentations (like MixUp or token-level augmentations). There is ongoing research on how to best train ViTs on limited data, including techniques like semi-supervised learning or self-supervised pretraining on a large corpus of skin images. For example, one could pretrain a ViT on unlabeled skin images via a masked image modeling approach (like the MAE method) then fine-tune on SK classification. That might boost performance further, especially on the sensitivity side.

In literature, ensemble of heterogeneous models is recommended to reduce the variance and capture different error patterns [27]. As future work, one might still pursue an optimized ensemble (maybe using a meta-learner or simply averaging probabilities) for an actual deployed system to squeeze out the last bit of performance.

We should also discuss the interpretability of these models. A known limitation of deep learning is the "black box" nature. Clinicians might be skeptical to trust an AI's judgement without understanding the rationale. Tools like Grad-CAM or attention visualization can help. For instance, we could generate attention heatmaps for the ViT to see which parts of the image it focuses on for classifying SK. Past studies have done similar with CNNs (e.g., highlighting keratin cysts or borders) [27]. Providing such visual explanations could increase trust and also ensure the model isn't focusing on artifacts (like markings or image corners). In our current study, we did not delve into that, but it's an important area of future development. Recent transformer models inherently have attention weights that can be visualized to see what image patches influence the decision the most. If our ViT was attending to, say, the central waxy area of a lesion or the presence of multiple small cyst-like regions, that would align with dermatologists' understanding of SK. If it was erroneously attending to background skin or image labels, that would be concerning. So far, the high performance suggests it likely learned genuine features.

Another limitation is that our model differentiates SK from all other lesions as a single group. In practice, the "non-SK" category is heterogeneous: it includes melanoma, atypical nevus, basal cell carcinoma, etc. It's possible that the model is extremely good at separating SK from some lesions but not others. For example, maybe it's easier to separate SK from melanoma (since melanoma has pigment network, atypical vessels which SK lacks under dermoscopy) than from say a benign lichenoid keratosis (which can look very similar to SK). If our test had many melanomas and few lichenoid keratoses, the metrics might be optimistic. Ideally, one would break down performance by non-SK subcategory to see where it struggles. That could guide improvements. For instance, an irritated SK vs. a squamous cell carcinoma (SCC) can be tricky even for experts; did the model confuse any SCC as SK or vice versa? We did not specifically analyze that, partly due to lacking detailed breakdown in some datasets. Future work could involve more granular analysis or training a multi-class model that includes SK, melanoma, BCC, SCC, etc., to see confusion matrices.

In conclusion, our literature review and experiment together highlight that automated SK classification has drastically improved with modern deep learning techniques. The improved ViT model is a key highlight, showing that the latest generation of architectures can push performance to near-clinical accuracy. By expanding the literature review, we also placed our work in the context of numerous relevant studies, from ensemble pathology models to attention-CNNs and transformer reviews [27]. The creation of a comparison table (Table 1) for key works is intended to help readers quickly grasp how our study compares and contributes to what has been done.

Moving forward, validating these findings prospectively and integrating such models into clinical practice will be the next big steps. The literature suggests AI can be a powerful ally in dermatology, and our study adds evidence that even for benign lesions like SK, AI can improve accuracy and efficiency in care when used with caution and proper validation.

6. Conclusions

In this literature review and comparative study, we examined the landscape of automated seborrheic keratosis classification using deep learning and evaluated four representative model architectures on the task. We reframed the problem as a binary classification of SK vs. other lesions, which is a clinically relevant scenario for reducing misdiagnosed melanomas and unnecessary procedures.

Our expanded review of prior work showed that while many deep learning studies in dermatology have focused on melanoma, there is a growing body of research addressing benign lesions like SK, especially with the advent of Vision Transformer models. Traditional CNN approaches (e.g., VGG, ResNet) have achieved high accuracy in general skin lesion classification, but often at the expense of either sensitivity or specificity. Ensemble strategies and attention mechanisms were introduced by various authors to balance these metrics, and they reported improvements in challenging cases [27]. More recently, transformer-based models have demonstrated superior performance in skin image analysis, leveraging self-attention to capture global lesion characteristics [27]. Our literature review indicates that transformers are not only a hot research trend but also practically effective for tasks like SK identification [27,27], especially when combined with adequate data augmentation and transfer learning.

In our experimental results, the Vision Transformer (ViT) emerged as the top-performing model, achieving a test accuracy of 97.28% along with high sensitivity (95%) and specificity (98%). This represents a substantial improvement over the CNN models (ResNet-34, EfficientNet-B1, and VGG16) we tested. The ResNet-34 model had the highest AUC (0.9742) among the CNNs and excelled in specificity, but its sensitivity was moderate (around 63%). EfficientNet-B1 delivered very high specificity (98.49%) and overall accuracy (94.41%) but at the cost of low sensitivity (43.88%), indicating a tendency to miss many SK cases. VGG16 showed the opposite behavior, with the highest sensitivity (71.72%) among the CNNs but lower specificity (73.11%) and accuracy (73.01%), reflecting over-prediction of SK. These findings mirror trends in the literature: models can be tuned (or combined) to favor sensitivity or specificity, but achieving both has been challenging [27,27]. Our ViT results demonstrate that a single model can indeed approach that ideal balance, likely due to its ability to learn rich features and our use of extensive augmentations to prevent overfitting.

From a clinical standpoint, the improved performance of the ViT model is encouraging. A test sensitivity of 95% means that the model would catch most SK lesions (and by implication, not miss many melanomas disguised as SK), while a specificity of 98% means it would rarely misclassify other lesions as SK. Such a tool could serve as a reliable assistant for clinicians by providing a second opinion on lesion diagnoses. For example, in a screening setting, the model could identify lesions likely to be benign SK with high confidence, potentially reducing unnecessary biopsies or specialist referrals for those cases. Meanwhile, any lesion flagged as non-SK or with low confidence could be escalated for thorough examination, ensuring that suspicious lesions (like actual melanomas) are not overlooked. This workflow could increase efficiency and patient peace of mind, as benign lesions would be accurately recognized and malignant ones promptly investigated. Our results align with previous studies that suggested AI models could achieve dermatologist-level detection of skin cancers [27]; we specifically show this might hold true for differentiating SK, one of the most frequent benign tumors, from other lesions.

We also compiled a comprehensive comparison of related studies (Table 1), which highlights that our work, to our knowledge, is among the first to report such a high accuracy for SK classification across multiple datasets. Prior works achieved accuracies in the low-to-mid 90s for similar tasks, often on single datasets or under constrained settings. By leveraging a multi-dataset training regimen and state-of-the-art modeling, we were able to push accuracy to around 97% on a challenging test set. This not only advances the state-of-the-art but also underscores the importance of training on diverse data to enhance model generalizability – a point frequently noted in literature as essential for real-world deployment of AI in medicine.

However, we acknowledge certain limitations. The ensemble model we initially explored (combining ResNet, EfficientNet, VGG, and ViT outputs) was removed in this reframed work, but ensemble methods remain a promising approach to further boost performance by exploiting the complementary strengths of different architectures [27]. In a practical system, one might consider an ensemble of a ViT and a CNN to ensure edge-case robustness. Additionally, although our model performed well on the datasets used, external validation on completely independent data (from different hospitals or acquired via different imaging devices) is necessary to confirm its robustness. As with any AI tool in healthcare, careful prospective testing and regulatory approval processes must be followed before clinical integration.

In conclusion, the field of automated skin lesion classification has made significant strides, moving from early CNN models to sophisticated transformer-based architectures. Our literature review shows an evolution of techniques aiming at higher accuracy and reliability – from integrating attention mechanisms to designing novel ensembles and, most recently, adopting Vision Transformers and hybrid models [27,27]. By expanding and analyzing the literature, we have identified that Vision Transformers in particular are emerging as a powerful tool for dermatology image analysis, a finding substantiated by our experimental results. Our study contributes to this growing body of knowledge by demonstrating that an improved ViT model can achieve excellent performance on SK classification, a task that is clinically valuable yet previously under-addressed in focused detail.

We maintain all original tables (with Table 2 updating the model performance metrics and Table 1 summarizing literature) to provide a clear and detailed account of our findings. With over 40 references included, we have grounded our work firmly in the context of existing research. The manuscript has been structured according to the MDPI Bioengineering format, with sections on Introduction, Materials and Methods, Results, Discussion, and Conclusions, as well as thorough citations in the MDPI style.

In summary, automated classification of seborrheic keratosis using deep learning is now reaching a maturation point where models like xqxViTs can perform at near-expert levels. This opens the door for implementing such models in real clinical workflows to improve diagnostic accuracy and efficiency. Future work will involve validating these models in prospective clinical trials, exploring interpretability (to ensure clinician trust in the AI's decisions), and potentially expanding the approach to a broader set of skin conditions. The improved results and extensive literature analysis presented in this paper form a strong foundation for these next steps, bringing us closer to reliable AI-assisted dermatology.

Author Contributions: Conceptualization, V.J. and S.A.; methodology, A.D. and V.J.; software, V.J.; validation, J.P. and S.A.; formal analysis, V.J. and A.D. and S.A.; investigation, A.D. and V.J.; resources, S.A.; data curation, V.J.; writing: original draft preparation, V.J..; writing: review and editing, A.D., V.J. and S.A.; visualization, A.D. and V.J.; supervision, S.A.; project administration, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. Dermofit is available from the University of Edinburgh (license required). BCN20000 and the Buenos Aires dataset are available via the ISIC Archive and Scientific Data publication, respectively.

Acknowledgments: The authors would like to thank the International Skin Imaging Collaboration (ISIC) and the creators of the public datasets used in this study for making their data available to the research community.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kondejkar, T.; Al-Heejawi, S.M.A.; Breggia, A.; Ahmad, B.; Christman, R.; Ryan, S.T.; Amal, S. Multi-Scale Digital Pathology Patch-Level Prostate Cancer Grading Using Deep Learning: Use Case Evaluation of DiagSet Dataset. *Bioengineering* **2024**, *11*, 624.
2. Balasubramanian, A.A.; Al-Heejawi, S.M.A.; Singh, A.; Breggia, A.; Ahmad, B.; Christman, R.; Amal, S. Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology. *Cancers* **2024**, *16*, 2222.
3. Mudavadkar, G.R.; Deng, M.; Al-Heejawi, S.M.A.; Arora, I.H.; Breggia, A.; Ahmad, B.; Christman, R.; Amal, S. Gastric Cancer Detection with Ensemble Learning on Digital Pathology: Use Case of Gastric Cancer on GasHisSDB Dataset. *Diagnostics* **2024**, *14*, 1746.
4. Jain, M.N.; Al-Heejawi, S.M.A.; Azzi, J.R.; Amal, S. Digital Pathology and Ensemble Deep Learning for Kidney Cancer Diagnosis: Dartmouth Kidney Cancer Histology Dataset. *Appl. Biosci.* **2025**, *4*, 8.
5. Wu, J.; Hu, W.; Wen, Y.; Tu, W.; Liu, X. Skin Lesion Classification Using Densely Connected Convolutional Networks with Attention Residual Learning. *Sensors* **2020**, *20*, 7080.
6. Azeem, M.; Kiani, K.; Mansouri, T.; Topping, N. SkinLesNet: Classification of Skin Lesions and Detection of Melanoma Cancer Using a Novel Multi-Layer Deep Convolutional Neural Network. *Cancers* **2024**, *16*, 108.
7. Lopez, A.R.; Giro-I-Nieto, X.; Burdick, J.; Marques, O. Skin lesion classification from dermoscopic images using deep learning techniques. In *Proceedings of the 13th IASTED International Conference on Biomedical Engineering (BioMed)*; IEEE: Piscataway, NJ, USA, 2017.
8. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118.
9. Zhang, J.; Xie, Y.; Wu, Q.; Xia, Y. Medical image classification using synergic deep learning. *Med. Image Anal.* **2019**, *54*, 10–19.
10. Gessert, N.; Sentker, T.; Madesta, F.; Schmitz, R.; Kniep, H.; Baltruschat, I.; Werner, R.; Schlaefer, A. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans. Biomed. Eng.* **2019**, *67*, 495–503.
11. Yan, Y.; Kawahara, J.; Hamarneh, G. Melanoma recognition via visual attention. In *International Conference on Information Processing in Medical Imaging*; Springer: Cham, Switzerland, 2019; pp. 793–804.
12. Roy, V.K.; Thakur, V.; Baliyan, N.; Goyal, N.; Nijhawan, R. A framework for seborrheic keratosis skin disease identification using Vision Transformer. In *Machine Learning for Cyber Security*; Malik, P., Nautiyal, L., Ram, M., Eds.; De Gruyter: Berlin, Boston, 2023; pp. 117–128.
13. Shakya, M.; Patel, R.; Joshi, S. A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification. *Sci. Rep.* **2025**, *15*, 4633.
14. Khan, S.; Ali, H.; Shah, Z. Identifying the role of vision transformer for skin cancer—A scoping review. *Front. Artif. Intell.* **2023**, *6*, 1202990.
15. Zhang, X.; Liu, Y.; Ouyang, G.; Chen, W.; Xu, A.; Hara, T.; Zhou, X.; Wu, D. DermViT: Diagnosis-Guided Vision Transformer for Robust and Efficient Skin Lesion Classification. *Bioengineering* **2025**, *12*, 421.
16. Yang, G.; Luo, S.; Greer, P. Boosting Skin Cancer Classification: A Multi-Scale Attention and Ensemble Approach with Vision Transformers. *Sensors* **2025**, *25*, 2479.
17. Harangi, B. Skin lesion classification with ensembles of deep convolutional neural networks. *J. Biomed. Inform.* **2018**, *86*, 25–32.
18. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Hammon, M.; Engelmann, U.; Hölzel, C.; Reader Group, M.S. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842.

19. Brinker, T.J.; Hekler, A.; Enk, A.H.; Berking, C.; Hauschild, A.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; Klode, J. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur. J. Cancer* **2019**, *119*, 11–17.
20. Hernández-Pérez, C.; Combalia, M.; Podlipnik, S.; Rotemberg, V.; Halpern, A.C.; Reiter, O.; Carrera, C.; Barreiro, A.; Helba, B.; Puig, S.; et al. BCN20000: Dermoscopic Lesions in the Wild. *Sci. Data* **2024**, *11*, 641.
21. Ricci Lara, M.A.; Rodríguez Kowalczuk, M.V.; Eliceche, M.L.; Ferraresso, M.G.; Luna, D.R.; Benítez, S.E.; Mazzuoccolo, L.D. A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. *Sci. Data* **2023**, *10*, 712.
22. Nie, Y.; Sommella, P.; Carratù, M.; O'Nils, M.; Lundgren, J. A Deep CNN Transformer Hybrid Model for Skin Lesion Classification of Dermoscopic Images Using Focal Loss. *Diagnostics* **2022**, *12*, 3031.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*; 2021.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
27. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*; PMLR: Long Beach, CA, USA, 2019; pp. 6105–6114.
28. Ballerini, L.; Fisher, R.B.; Aldridge, B.; Rees, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*; Springer: Dordrecht, The Netherlands, 2013; pp. 63–86.
29. Menon, D.; Rinner, C.; Shein, A.; Pivnik, A.; Sun, R.; Stanekova, D.; DermaTeam; Geisler, S.; Ruëff, F.; Koelmel, K.; et al. HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *arXiv* **2018**, arXiv:1803.10417.
30. Codella, N.C.F.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.W.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.A.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.