OPEN ACCESS

ISSN: 2835-9496

Herrick G, et al. (2024): 454
DOI: 10.47991/2835-9496/AJCMR-138

Review Article

American Journal of Clinical and Medical Research

Integrating Deep Learning with Dermoscopic Imaging for Enhanced Melanoma Detection

Grace Herrick, BA[1], Mahnoor Mukarram, MS[2], Vivian Li, DO, MMS[3], Eliza Skemp, BS[2], Rebekah Mounger, BS[2], Kelly Frasier, DO, MS[3*], Haily Fritts, BS[4], Jordan Saunooke, BS[4]

[1]Alabama College of Osteopathic Medicine, Dothan, AL
[2]Midwestern University Arizona College of Osteopathic Medicine, Glendale, AZ
[3]Nuvance Health/Vassar Brothers Medical Center, Poughkeepsie, NY
[4]Idaho College of Osteopathic Medicine, Meridian, ID

***Corresponding author:** Kelly Frasier, Nuvance Health/Vassar Brothers Medical Center, Poughkeepsie, NY. Email: kellymarie frasier@gmail.com

Citation: Herrick G, Mukarram M, Li V, Skemp E, Mounger R, et al. (2024) Integrating Deep Learning with Dermoscopic Imaging for Enhanced Melanoma Detection. Ameri J Clin Med Re: AJCMR-138.

Received Date: 13 June, 2024; **Accepted Date:** 18 June, 2024; **Published Date:** 24 June, 2024

Abstract

The integration of deep learning with dermoscopic imaging presents a promising advancement in the early detection and diagnosis of melanoma, a deadly form of skin cancer. This review synthesizes the current literature on the application of AI-driven deep learning algorithms to dermoscopic images, highlighting significant improvements in detection accuracy and diagnostic efficiency. Key areas of focus include the technical intricacies of model training, emphasizing the critical role of diverse and extensive datasets to enhance algorithm robustness and generalizability. The review also addresses the challenges inherent in the interpretability of AI decisions, which is crucial for clinical acceptance and trust. Additionally, the potential of these technologies to reduce diagnostic errors and improve patient outcomes is examined. The integration of deep learning systems into clinical workflows is discussed, considering the operational and ethical implications. Future research directions are identified, such as the development of more transparent AI models, the creation of standardized evaluation metrics, and the exploration of hybrid models combining deep learning with traditional diagnostic methods. By providing a comprehensive analysis of these aspects, this review aims to guide future research and facilitate the adoption of deep learning technologies in clinical dermatology for enhanced melanoma detection.

Keywords: *Dermoscopic Imaging, Melanoma Detection, Artificial Intelligence, Diagnostic Accuracy, Clinical Integration, Dataset Diversity, Model Interpretability, Hybrid Models, Dermatology.*

Introduction

Melanoma is an aggressive form of skin cancer characterized by the malignant transformation of melanocytes. While it constitutes only about 1% of skin cancer cases, it is the leading cause of skin cancer-related deaths [1]. This discrepancy highlights the need for better preventative measures and diagnostic strategies. In 2022, the United States recorded approximately 100,000 new cases and 7,650 deaths due to invasive melanoma, revealing the substantial impact of this disease on public health [1].

Melanoma is particularly prevalent among older adults with lighter skin, individuals with a history of extensive sun exposure, and those with a genetic predisposition to the disease. However, it does not exclusively affect this demographic; individuals of all ages and skin types can develop melanoma, often with more challenging detection and outcomes in those with darker skin tones. Unlike in lighter-skinned individuals, melanoma in skin of color (SOC) is not as commonly linked to sun exposure and often manifests in sun-protected areas like the palms and soles, where melanoma is less suspected and often diagnosed at later stages [1]. This highlights the importance of considering diverse clinical presentations and risk factors when screening for melanoma across different population groups.

The integration of deep learning with dermoscopic imaging presents substantial opportunities for advancements in melanoma detection and diagnosis. Dermoscopic imaging, which allows for a detailed examination of skin lesions that are not visible to the naked eye, is instrumental in identifying melanoma at its earliest and most treatable stages. By incorporating deep learning algorithms, these images can be analyzed with greater precision and consistency, reducing the subjectivity associated with human analysis [2]. This also holds the potential to revolutionize the speed and reliability with which melanoma is identified, thereby improving outcomes and reducing mortality rates associated with the disease.

This narrative literature review aims to comprehensively explore the latest advancements in the diagnosis and management of melanoma, with a particular focus on the integration of deep learning with dermoscopic imaging. By examining the effectiveness of these technologies and strategies across diverse demographic groups, the review highlights critical areas for future research and potential improvements in clinical practices. Additionally, it assesses the impact of technological innovations on the accuracy of melanoma detection, especially in early stages where intervention is most effective. The review explores the challenges and successes of implementing these technologies in real-world settings, considers the ethical implications of automated diagnostics, and

discusses the potential for these tools to reduce disparities in healthcare outcomes. Ultimately, the review aims to provide a detailed analysis that informs both current and future approaches to deep learning in melanoma detection, allowing for more personalized, effective, and equitable treatment options.

Role of Dermoscopy in Melanoma Diagnosis

Dermoscopy is a diagnostic technique that utilizes a dermatoscope to closely examine and evaluate suspicious lesions. This method is a critical tool in enhancing the accuracy of skin diagnoses and effectively differentiating between melanomas, dysplastic lesions, and other skin cancers (e.g., basal cell carcinoma, squamous cell carcinoma) [3]. Through magnification and illumination, dermoscopy allows clinicians to observe subsurface skin structures and patterns that are otherwise not visible to the naked eye. Established dermoscopic criteria and key pathological features, such as the ABCDE's of melanoma, provide guidance for clinicians in determining possible malignancy of skin lesions [4]. Given its non-invasive nature and ability to provide more detailed visualizations than the naked eye, the use of dermatoscopes has expanded beyond the evaluation of skin lesions. This versatile tool is now frequently used to examine a variety of dermatological conditions affecting the skin, hair, scalp, and nails.

Challenges in Early Detection

Although dermoscopy has revolutionized skin checks, it still has its limitations. There is inter-observer variability, where different clinicians might interpret the same dermoscopic images differently, leading to inconsistent diagnoses for the same lesions. The variability highlights the inherent subjectivity of dermoscopic analysis and underscores the need for more objective and reliable diagnostic methods. Such advancements would greatly enhance early melanoma detection and reduce diagnostic discrepancies in clinical practice.

Detecting melanoma, especially in its early stages, poses significant challenges. These are primarily due to its clinical presentation, which often resembles benign lesions that closely mimic the appearance of melanoma, which can complicate accurate diagnosis and timely intervention. Benign mimickers, such as dysplastic nevi (DN), share visual characteristics with melanoma, making it difficult for clinicians to distinguish them solely through visual inspection [5,6]. Their appearance frequently overlaps with the ABCDE-melanoma detection criteria, commonly presenting with ill defined borders, variable coloring, bumpy surfaces, and larger in size than common nevi [7,8]. While these lesions are often diagnosed based on their clinical appearance, they are occasionally biopsied to rule out melanoma.

Histologically, DN present with an irregular appearance similar to that of melanoma due to its dysplastic growth pattern [9,10]. While these lesions are benign, they serve as important indicators, identifying patients who are at a greater risk of developing melanoma in the future [8,9]. As such, for patients who present with DN or are diagnosed with dysplastic nevi syndrome (DNS), routine skin checks are essential. These regular examinations play a crucial role in the early detection of melanoma, enhancing the chances for effective treatment and better outcomes.

The variability in lesion appearance across different individuals adds another layer of complexity, increasing the risk of misdiagnoses or delays in detecting malignant changes. The limitations of visual inspection for early melanoma detection are profound, as many melanomas do not exhibit distinct or recognizable features until they have progressed to more advanced stages. Balancing the identification of early-stage melanoma without over unnecessarily biopsying benign lesions remains a challenge [11]. Such difficulties hinder effective early intervention strategies, ultimately increasing the risk of poor outcomes for patients.

Deep Learning in Dermoscopic Analysis

The integration of deep learning, a subset of artificial intelligence (AI), into dermoscopic analysis offers promising advancements in the detection and diagnosis of melanoma, eliminating clinician-diagnostic variability, misdiagnosis, and delayed treatment. Deep learning algorithms are capable of performing automated feature extraction from dermoscopic images, significantly improving diagnostic accuracy and consistency [2]. These techniques leverage large datasets and sophisticated neural networks to identify patterns and features associated with disease pathology, which may not be immediately apparent to human observers [12]. By providing a standardized approach to image analysis, deep learning has the potential to reduce inter-observer variability and subjective interpretation, ensuring more uniform and reliable diagnostic outcomes. The improved accuracy and consistency offered by deep learning approaches represent a crucial step forward in the early detection of melanoma, paving the way for more effective and timely treatments.

Understanding Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have emerged as a transformative tool in the realm of image analysis. Their ability to adaptively learn spatial hierarchies of features from inputted images has revolutionized how visual data are processed and interpreted. The core of CNNs has several key architectural components, known as layers [13]. These layers begin with an input layer that receives the input image and ends with an output layer that provides the predictive factors of the image. A few important layers to discuss in detail include convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply a series of learnable filters to the input image, which detect specific filtered or selected patterns within the image [13]. As the filters of convolutional layers move across the image, this feature is able to map out and capture hierarchies such as textures, edges, and shapes.

Subsequently, pooling layers work to reduce the spatial dimensions from the output of convolutional layers. Pooling layers decrease the width and height of a map while maintaining the same depth to simplify the original image model [14]. This is typically executed via max pooling or average pooling. Max pooling retains the most important features of an image while reducing its size, while average pooling smooths out the image features while also reducing its size; these techniques help to decrease the number of parameters within the network, ensuring that the image is less computationally intensive to analyze [14]. Faster processing times are a key benefit of this efficient method. By down-sizing the image, the network becomes less sensitive to small changes in the input, which offers several advantages including improved generalization, stability, and reduced overfitting [15].

Lastly, fully connected layers, or dense layers, connect every neuron in each layer to every neuron in the subsequent layer.

Fully connected layers are typically situated towards the end of the network and are responsible for high-level reasoning and classification based on the features extracted by the preceding convolutional and pooling layers [16]. This layer integrates and interprets learned features to classify an image into a specific category and then make predictions about the image, such as categorizing lesions as benign or malignant. With the high precision layering technique that CNNs employ, the likelihood of false positives and false negatives in melanoma diagnosis can be significantly reduced [17]. Furthermore, CNNs can serve as a second opinion, increasing diagnostic confidence and enabling informed decision-making practices. With these early and accurate melanoma diagnoses, there are timely interventions which can prove to better patient survival rates and outcomes.

One study assessing deep learning algorithms across various datasets for 12 different skin diseases, including melanoma, determined the area under the curve (AUC) for melanoma predictions to be 0.96 ± 0.00 within one dataset [18]. This suggests almost no variability in the model's performance and indicates excellent diagnostic accuracy. Although the study did find that the melanoma specificities differed between two datasets, potentially due to skin coloration outside of the lesions, a solution was presented to generate different models for those with SOC. In 2021, an evaluation of 19 studies comparing the performance of AI models for automated melanoma classification to human experts showed that deep integrated learning CNN's performed superiorly or at least equivalently to clinicians [20].

In addition to these foundational components, transfer learning has become a powerful approach in the application of CNNs to dermoscopic image analysis. Transfer learning leverages models that have already been trained on large and diverse datasets to initialize the network, which can then be fine-tuned on a specific task, such as melanoma detection [21,22]. For instance, models pre-trained on the ImageNet, a database with millions of annotated images across thousands of categories, have also proven to be highly effective [17]. Because these pre-trained models are already capable of identifying a variety of features and patterns across a wide range of images, detection is significantly accelerated. The model only needs to adjust its weights to specialize the specific features relevant to melanoma [17,22]. This fine-tuning process entails further training the model on a smaller, more finite, and specific dataset related to melanoma. By leveraging these pre-trained models, researchers and clinicians can develop highly reliable tools for early and precise melanoma diagnosis.

Model Training Strategies

Effective training strategies are crucial for maximizing the performance of CNNs in dermoscopic image analysis for melanoma detection. One such strategy is data augmentation, which enhances the diversity of the training dataset without the need for additional data collection. This technique is vital in preventing overfitting, in which a model performs well on training data but poorly on unseen and new data, while also improving the model's generalization capabilities [15]. Common data augmentation methods include geometric transformations of the input images, such as cropping, flipping, rotating, and scaling. Altering of the geometry of images helps the model learn invariance to these transformations; in other words, the object's orientation or size should not affect its identification, so melanoma should be able to be detected regardless of its position in the image [23]. Color augmentations, which adjust the brightness, contrast, saturation, and hue, allow the model to become robust to variations in lighting conditions and color distributions [23]. This is particularly important as no two individuals share the exact same skin tone. A study on the symmetry of pigmented lesions analyzed their geometry, texture, and color between benign lesions and melanoma [24]. Using this symmetry analysis for skin cancer diagnosis, the study achieved a sensitivity and accuracy rate of 78% and 72%, respectively.

Additionally, Generative Adversarial Networks (GANs) can produce synthetic images that closely resemble the original dataset, further enriching the training data. GANs that have learned melanoma detection can then create new images of melanoma to serve as additional examples to learn from without having to manually collect more data. Qin et al. proposed their own GAN for skin lesions and compared their model to other GAN models [25]. The proposed skin lesion style-based GAN was evaluated against the International Skin Imaging Collaboration (ISIC) dataset in 2018, and accuracy, sensitivity, specificity, average position, and balanced multiclass accuracy metrics were accounted for. Increases of 1.6% raised accuracy to 95.2%, increases of 24.4% raised sensitivity to 83.2%, increases of 3.6% raised specificity to 74.3%, increases of 23.2% raised average precision to 96.6% and increases of 5.6% raised balanced multiclass accuracy to 83.1%. Therefore, adding synthesized images created from GANs to the training set model significantly improved the model's performance in correctly identifying both malignant and benign lesions, as well as in maintaining a high level of precision and balanced accuracy across multiple classes.

Another critical aspect of model training is the choice of loss functions and optimization algorithms. Loss functions quantify the difference between the predicted and actual outputs, guiding the model's learning process. Cross-entropy loss is commonly used for classification tasks, as it measures how well the predicted probabilities match the actual class labels [26]. Cross-entropy would specifically be employed to categorize benign versus malignant melanoma. Focal loss, on the other hand, is designed to address class imbalance by down-weighting the loss assigned to well-classified examples, thereby focusing more on hard-to-classify instances [26]. Optimization algorithms, such as Stochastic Gradient Descent (SGD), update the model parameters based on a random subset of data [27]. This gradually improves accuracy of analyzing networks in small steps and with each iteration, leading to faster convergence and reduced memory usage.

Datasets for Dermoscopic Image Analysis

The use of high-quality datasets is indispensable for training and evaluating deep learning models in dermoscopic image analysis. These datasets provide a diverse range of images and annotations, which are crucial for developing both robust and accurate models. Several key datasets have been instrumental in advancing research in this field. The ISIC Archive is a large repository of dermoscopic images that features various skin conditions along with annotations made by experts [23]. Each image is labeled precisely with information about the condition it represents, which is crucial to effectively train models. The ISIC Melanoma Project, a subset of the ISIC Archive, focuses specifically on melanoma detection, offering a comprehensive dataset for developing and benchmarking models that aim to

identify melanoma. The BCN20000 dataset also facilitates research in skin lesion analysis as it comprises dermoscopic images from the Hospital Clínic of Barcelona. Similarly, the Human Against Machine (HAM) 10000 dataset provides a large collection of multi-source dermoscopic images of common pigmented skin lesions from different sources and conditions, serving as another valuable resource for training deep learning models [28]. A study that used the HAM10000 and BCN20000 datasets combined with an algorithm known as artificial jellyfish (AJS) and the Feature-based Optimized Weighted Feature Set (FOWFS) strategy demonstrated both accuracy and precision in skin lesion diagnosis [29].

Preparing high-quality training data involves several critical steps, including data annotation and preprocessing. Arguably the most important aspect is the expert annotation, which ensures that the labels used in the training process are accurate and reliable. Data cleaning is also a vital process, which involves removing irrelevant images that could negatively impact the model's performance [30]. Normalization, the process of scaling the pixel values to a common range, helps in stabilizing and accelerating the training process [30]. These datasets create a library of images, complete with expert notes, that teach CNNs to accurately recognize and diagnose various skin conditions. By cleaning up the images and arranging them into similar formats, the training process smoothens, and resulting models become increasingly reliable for practices such as melanoma detection. Thus, applying CNNs to dermoscopic image analysis involves understanding and implementing various architectural components, leveraging transfer learning, adopting effective model training strategies, and utilizing high-quality datasets. These elements collectively contribute to the development of accurate and reliable models for melanoma detection and other dermatological tasks.

Performance Evaluation and Benchmarking

Sensitivity, specificity, and area under the curve (AUC)

Critical metrics in the evaluation of melanoma detection include sensitivity, specificity, and AUC. Sensitivity, also known as the true positive rate, measures the proportion of actual melanoma cases correctly identified by a diagnostic tool. High sensitivity is crucial for minimizing false negatives and thus ensuring that patients with melanoma receive timely treatment. Specificity, or the true negative rate, measures the proportion of non-melanoma cases accurately identified, therefore reducing the number of false positives and potentially unnecessary medical treatments. The AUC is derived from the receiver operating characteristic (ROC) curve, which plots sensitivity against 1-specificity. An AUC of 1 represents a perfect test, while an AUC of 0.5 indicates a test with no discriminative power. In the context of melanoma detection, a higher AUC value signifies better overall diagnostic performance. The AUC is a valuable measure as it reflects the model's ability to distinguish between positive and negative cases across various threshold settings [31]. Literature has supported the efficacy of deep learning models in achieving high sensitivity and specificity. For example, a CNN developed to classify dermatological lesions demonstrated a sensitivity of 72.1% and a specificity of 91.0%, with an AUC of 0.91, surpassing the performance of dermatologists (AUC 0.87) [32]. This study highlights the potential of AI in augmenting diagnostic accuracy and dependability in dermatologic practice.

Diagnostic accuracy of human experts compared to Deep Learning and traditional dermoscopic evaluation

The diagnostic accuracy of human experts in melanoma detection using dermoscopy has traditionally served as a benchmark in clinical practice. Dermoscopic evaluation relies on the visual assessment of skin lesions, requiring substantial expertise and experience. Studies have indicated that experienced dermatologists have a high diagnostic accuracy, but this can vary significantly. For instance, Vestergaard et al. reported a diagnostic accuracy ranging from 75% to 85% among dermatologists using dermoscopy [33].

Research comparing the performance of a combined CNN model with human medical personnel, including 62 board-certified dermatologists, in expert-level diagnosis of nonpigmented skin cancer revealed that the AUC-ROC of the trained CNN was higher than human ratings [34]. The sensitivity of the CNN was also higher than that of human raters. Moreover, the study demonstrated that the CNN had a greater percentage of correct specific diagnoses compared to human raters, however, the percentage correct was less than that of experts. These results support that neural networks are capable of classifying dermoscopic images with significant accuracy and have the potential to contribute greatly to healthcare environments.

Inter-observer variability among clinicians

Interobserver variability refers to the differences in diagnostic decisions made by different clinicians when evaluating the same patient. This variability can significantly affect the reliability of melanoma detection using traditional dermoscopy. Vestergaard et al. highlighted that the level of agreement among dermatologists in diagnosing melanoma can range from moderate to substantial, indicating variability in diagnostic consistency [33]. Hence, the introduction of AI and deep learning models into clinical practice holds the potential to reduce inter-observer variability. Haenssle et al. concluded that CNN reduced interobserver variability among dermatologists by providing a consistent second opinion, which helped standardize diagnostic decisions [35]. AI and deep learning models provide consistent results determined by standardized algorithms, thereby minimizing the subjective nature of human diagnosis.

Deep learning also holds potential as a diagnostic tool distinct from traditional machine learning approaches. The goal of implementing deep learning is to develop accurate diagnostic techniques capable of analyzing raw data without requiring human input or oversight. This is particularly beneficial in rural areas lacking access to medical imaging experts, as it provides an automated system capable of disease detection [36]. The ability to rely on AI for diagnosis with the same accuracy as healthcare professionals represents a significant advancement in global healthcare. Moreover, traditional diagnostic imaging is often hindered by time-consuming analysis and susceptibility to human error [37]. Deep learning addresses these issues by delivering timely and accurate results, free from the effects of human fatigue.

Discussion of Different Deep Learning Models

To date, numerous studies have shown the effectiveness of deep learning models to identify melanoma and other skin lesions through dermoscopic images. These studies have used various data sets and learning models, showing mixed results. It is crucial to recognize differences in performance to determine the

models best suited for clinical practice. Several of the primary learning models are discussed here.

Inception

Inception is one of the primary models used in identifying melanoma from dermoscopic images. Leveraging its intricate architecture, Inception-based models excel at capturing specific details and patterns in dermoscopic images that indicate melanoma. Work by Haenssle et al. showed that Inception-based technology was able to outperform dermatologists in identifying melanoma with an AUC of 0.86 compared to 0.79 in the dermatologist group [35]. Additionally, Esteva et al. highlighted the capability of deep neural networks to perform on par with board-certified dermatologists in diagnosing skin cancer, including melanoma, from clinical images [32]. The model in this study utilized 129,450 images and achieved an AUC of 0.91 for melanoma detection, similar to that of board-certified dermatologists. These findings underscore the potential of Inception-based deep learning methods to detect and diagnose melanoma from dermoscope images.

ResNet

ResNet-based models have also demonstrated strong performance in melanoma detection in several studies. A study focusing on the classification of benign and malignant lesions using a ResNet152 structure demonstrated that the deep learning algorithm achieved high sensitivity (82%) and specificity (92.5%) in detecting melanoma, with a reported accuracy of 90% [38]. Additionally, Han et al. used a ResNet CNN model to identify 12 skin diseases, including melanoma, with notable success. This model was trained on a dataset of 129,450 clinical images and achieved a high level of performance comparable to that of board-certified dermatologists [18]. Specifically, the CNN model achieved a sensitivity of 91.0±4.3% and a specificity of 90.4±4.5% for melanoma detection, demonstrating its effectiveness in identifying malignant skin lesions. This model's ability to diagnose across 12 different skin diseases, rather than focusing solely on melanoma, may enhance its applicability in diverse clinical settings.

Inception-ResNet

The Inception-ResNet technology takes the existing strengths of Inception and adds residual connections to enhance accuracy [39]. Work using this model shows both improvements in accuracy and time to detection, a crucial element when considering the practicality of these tools. Singh et al. produced an accuracy of 96% using the ISIC 2020 dataset with a detection time of 39 seconds [40]. Using a similar model with the HAM10000 dataset, Alwakid et al. achieved an accuracy of 91% [41]. Both of these studies used standardized datasets to train their deep learning tool, showing improved accuracy compared to prior models.

EfficientNet

As a newer CNN architecture, EfficientNet has not been as widely studied in the detection of melanoma. However, a recent study using ISIC-2019 and ISIC-2020 datasets achieved an AUC of 0.97, outperforming other models in diagnosing melanoma [42]. EfficientNet offers unique benefits, as it is highly scalable and provides a systematic approach to expanding in three dimensions: depth, width, and resolution [43]. As technology progresses, such as with EfficientNet, available tools to detect melanoma should perform more accurately and efficiently, making them more applicable to clinical practice.

Challenges in Performance Evaluation and Benchmarking

While accuracy continues to improve within deep learning models, significant challenges still exist in dataset diversity and standardized evaluation protocols. These areas must be addressed to produce robust technology that can be used globally in medical practice. Dermoscopic image datasets used for training and evaluation in melanoma detection often suffer from biases and lack diversity, which can impact the generalizability of deep learning models. Biases in datasets, such as overrepresentation of light-skinned individuals, can lead to model performance disparities and hinder real-world applicability. Guo et al. have pointed out significant disparities that exist within current research on AI applications in dermatology [44]. In their review of 136 studies, they found that only six studies disclosed the skin types of the image sources. Among these, only two studies included type VI skin, with a total of just five subjects across the studies. It is crucial to address these biases, as melanoma tends to be more deadly among those with SOC, likely due to later stages of diagnosis [45]. Building deep learning models that can accurately diagnose melanoma in SOC may alleviate health disparities seen in dermatology.

In a multicenter observational study conducted by Mitre et al., the accuracy of an AI algorithm from the ISIC 2020 grand challenge was evaluated in a cohort of 100 non-Hispanic Black individuals [46]. The study revealed significant inaccuracies: 95.7% of benign volar lesions, 98.6% of benign dorsal skin lesions, and 100% of benign nail lesions were incorrectly identified as melanoma, resulting in specificities of 4.3%, 1.4%, and 0%, respectively. These results contrast with those of a prior study by Marchetti et al., which reported a sensitivity of 96.8% and specificity of 37.4% using the same AI algorithm but tested primarily on a cohort that was 96% white [47]. This discrepancy underscores the critical issues of accuracy and bias that arise when AI algorithms are trained on datasets that do not adequately represent the diversity of the population they serve.

Furthermore, the absence of standardized evaluation protocols presents a significant challenge in the objective assessment and comparison of deep learning models for melanoma detection. Without uniform benchmarking standards, it becomes challenging to determine the relative performance of different models and methodologies accurately. Addressing this requires the development of standardized evaluation metrics, methodologies, and datasets for consistent performance assessment across studies. Prior deep learning models have used a variety of datasets to train their models. For example, Haenssle et al. used 300 images selected from the University of Heidelberg image library [35] More recent research, however, has shifted towards using standardized datasets, such as ISIC and HAM10000 [40,41]. The International Skin Imaging Collaboration Melanoma Project has been addressing these shortcomings by organizing annual computer science challenges and providing standard datasets [48]. These initiatives facilitate standardized comparisons between learning models, while also promoting collaboration and innovation within the field.

Interpretability and Explainable AI

Importance of Interpretability in Clinical Decision Support

A significant obstacle to physicians accepting and integrating the use of AI within the clinical setting is the interpretability and explainability of the AI system. CNN is often referred to as "black box" technology due to the current limitations in

understanding how data is integrated to formulate a result and what defining variables the generated outcome is founded upon [49]. In order for a physician to confidently agree or disagree with the AI system, the ability to detect where an error in reasoning has occurred, whether that be within the physician's own thought process or the AI system, must be discernible. Explainability of CNN provides the foundation for experts within a field to trust outputs by AI systems and to feel comfortable incorporating its recommendations into clinical decision making.

The current lack of interpretability of CNN raises concerns on its ability to comply with ethical and regulatory guidelines of medical practice. One concern is the ability to maintain patient autonomy in regards to consent for use of personal information within the AI algorithm. Currently, there is a lack of consensus on how patient consent should be obtained or if it is even necessary to obtain at all [50]. This could have detrimental effects on the patient-doctor relationship and undermines the patient's autonomy if the use of an AI system or patient information occurs without explicit consent. If physicians are unable to effectively communicate the reasoning for their professional recommendation formulated with the aid of CNN, it becomes impossible for the patient to have an active role in their care and further decreases their autonomy. For patients and physicians to confidently follow the guidance of an AI decision, it is crucial to understand how that decision was made and to ensure that the variables deemed most important by the AI align with the patient's priorities. For example, an AI system that is taught to prioritize patient survival may not provide decision making that is suitable for a palliative care patient whose priority is to reduce their suffering [50].

Another concern is the extent of liability that falls on physicians that utilize AI systems for clinical decision making. *Would a physician be liable if they did not inform the patient of the risk and benefits of utilizing the AI system? What are the implications of using an AI system whose reasoning is not explainable in the case of a poor medical outcome, and does this give grounds for liability for medical malpractice*? Consent, privacy, standardizing use of data, and liability are topics of ongoing debate with accountability of AI still in its early stages of development. The explainability of AI may ultimately become a requirement for these systems under data protection law. Despite dermatologists reporting concerns for inaccuracies and risk of lawsuits, an overwhelming majority see AI as a promising addition to the field that will ultimately improve patient care once optimized [51,52].

Techniques for Enhancing Interpretability

For clinicians to comfortably integrate AI systems into clinical practice, the AI algorithms must be transparent and understandable to users. Recently, different types of explainability artificial intelligence (XAI) models have been a topic of research and discussion in attempts to improve transparency of AI decision making. Many of these techniques are considered post-hoc interpretation techniques, as they attempt to analyze the neural network after it has been trained. One proposed method is the application of saliency/attribution maps. Saliency maps offer valuable insights by highlighting the pixels within an image that the AI algorithm has identified as most important, illustrating what information guided the network's decision-making process [53]. While this method allows users to see what part of the image the AI system determines influential, it does not provide further insight into how this is incorporated into neural network decisions and is difficult to interpret into meaningful and easily understandable information. Another limitation of gradient-based techniques like saliency maps is the potential for inputs to become saturated, which ultimately diminishes the importance assigned to what would be considered a relevant area [54].

Another method is the perturbation-based technique that applies discrete modifications to each variable to measure its contribution to the outcome [55]. An example of this is Local Interpretable Model-Agnostic Explanations (LIME), which provides an interpretation of an outcome from the original model by taking predictions of the AI model and approximating a simpler version that is subsequently used to interpret the original outcome [56]. However, a drawback to this technique is that it analyzes an explanation for a single point that may not be generalizable across different inputs. This limitation underscores the need for careful application and possibly supplementary methods to ensure broader relevance and applicability.

While local methods like LIME produce an explanation for a single point within a set of points, Concept Activation Vector (CAV) is a type of global perturbation method that provides an explanation for the entire set of points by training a linear classifier to separate concepts from random images [57]. Testing Concept Activation Vector (TCAV) provides a score to indicate the importance of each concept towards creating the prediction. In a study by Kim et al., physicians using the TCAV method for predicting diabetic retinopathy (DR) were able to identify which concepts the AI system emphasized [57]. This insight helped them determine which variables the AI deemed most important for diagnosing DR, increasing their understanding of the AI-assisted diagnostic process. Concept-driven explainability of the AI model provides opportunities for physicians to collaborate with the CNN model and incorporate their professional expertise to maximize the overall capabilities of AI modalities.

Ethical Considerations

The adoption and integration of electronic health records (EHRs) has led to the accumulation of massive amounts of patient data. One significant challenge associated with this large volume of data is the ability to navigate the information effectively to facilitate clinical decision-making and improve patient care [58]. The advent of AI technology, capable of performing thorough analyses of secondary patient data, offers numerous advantages over traditional primary data collection methods. A principal argument for the integration of AI algorithms for analysis of EHRs is their ability to analyze patient data through longitudinal assessments at a fraction of the cost associated with primary data collection. AI technology can reuse existing patient data, eliminating the need for additional patient recruitment and streamlining the longitudinal processes necessary for effective research.

Informed consent is a principle ensuring that a fully aware and competent patient intentionally permits healthcare professionals to use their information. This process is characterized by clear transparency in the informational exchange between the physician and the patient, which is critical for the ethical evaluation of such transactions. However, the use of AI technology is viewed by some as a breach of this transparency, presenting obstacles to the use of EHR AI analysis for secondary

studies [59]. This controversy over transparency segues into broader discussions about the implications of AI integration in healthcare settings.

The question of whether patients should be informed about the use of AI in decision-making remains a significant topic of debate. The consensus is that when AI serves solely as a support tool in decision-making, obtaining patient consent might not be necessary. However, when AI plays a decisive role in determining the diagnosis or course of treatment, transparency becomes imperative, and patients should be fully informed. A pertinent issue in this debate is whether consent needs to be re-obtained if the patient's information is used in previously unidentified platforms [59].

Key reasons to advocate for patient awareness of AI involvement in their healthcare include the risks of cyberattacks and data breaches, potential systematic biases in the algorithms, and the possibility of mismatches between AI assumptions and individual patient backgrounds [60]. Although the integration of AI with patient health records does pose certain risks, the establishment of clear and universal guidelines can protect patient privacy while maximizing the benefits derived from extensive, pre-existing data collections. Governing the use of AI in healthcare is complex, but the long-term advantages of deep learning, such as the ability to sift through millions of medical records for pertinent information and pattern recognition, are believed to substantially outweigh the drawbacks [59]. This capability not only enhances the efficiency of medical diagnoses but also significantly reduces the time and effort required compared to manual analysis by human experts.

Future Directions

Addressing Dataset Biases and Diversity

The ISIC, one of the largest and most frequently used databases, primarily consists of data from fair-skinned populations in the United States, Europe, and Australia [61]. The success of deep learning algorithms heavily depends on both the volume and quality of data that they learn from. The apparent lack of diversity within these training datasets can introduce significant biases in the AI algorithms, limiting their applicability across diverse patient populations. This includes communities of color or individuals with rare diagnoses that are underrepresented in the data. A solution to this is to diversify image training sets to include a broader spectrum of melanoma and skin conditions prevalent among SOC, anatomical sites typical of minority populations, and rare atypical presentations of skin diseases [61].

To address dataset imbalances, techniques such as the Synthetic Minority Over-Sampling Technique and Adaptive Synthetic Sampling can be used. These methods generate synthetic examples of minority populations to enhance AI model performance in underrepresented groups [62]. While this strategy can help balance the dataset, it is important to understand that this can unintentionally lead to incorporation of additional biases into the dataset. Additionally, concept-based explanations like TCAV and others can be instrumental in identifying and understanding biases present in databases. These explanations offer insights that can guide the customization of network training to control which features are emphasized in the AI model.

Exploration of Hybrid Approaches and Multimodal Data Fusion

The use of dermatoscopic images in AI models has been shown to provide more accurate outputs than macroscopic images of the same lesion [63]. This is likely due to the dermatoscope's hardware, which introduces physically limiting variables like image size, lighting, and distance. However, recent research has been exploring the potential benefits of a multimodal approach to further optimize the accuracy of neural networks. Similar to how physicians utilize information from various sources to form a differential diagnosis, CNNs can also benefit from integrating data from multiple sources.

A study by Binder et al. demonstrated a significant improvement in a neural network's ability to differentiate early melanoma from benign pigmented lesions when trained on a combination of morphometric and clinical features, such as patient age and anatomic location, compared to morphometric features alone [64] Thus, future research should focus on effective data fusion occurring across different domains to include various types of images and textual inputs. Moreover, there is a growing trend in the development of hybrid neural networks that integrate multiple CNN features to improve the AI system's ability to discriminate between benign and malignant lesions [65]. Such advancements should be further investigated to maximize the potential of hybrid and multimodal AI systems in increasing the performance and reliability of CNN models in clinical settings.

Standardization of Evaluation Metrics and Benchmarking

The development and deployment of deep learning models for dermoscopic image analysis face significant challenges, particularly in the standardization of evaluation metrics and benchmarking protocols. To ensure consistent and reliable performance assessments, there is a pressing need for the standardization of these evaluation protocols, and for benchmarking datasets to be made publicly available. Without standardization, it becomes difficult to compare the effectiveness of different models, hindering the progress and validation of new techniques.

Currently, the lack of uniform evaluation criteria means that researchers use different metrics and datasets to test their models, which introduces variability in reported performances. To facilitate fair and meaningful comparisons, it is essential to develop standardized evaluation protocols that provide a common framework for assessing models. These protocols should clearly define guidelines for key metrics such as sensitivity, specificity, accuracy, and the AUC. The ISIC Archive serves as a perfect example of a publicly available benchmarking dataset that can act as a reference point for evaluating new models. Making such resources accessible enables researchers to effectively build upon each others' work, fostering collaboration and accelerating advancements in the field.

Addressing ethical concerns in AI deployment

Federated learning involves training machine learning models across multiple servers that hold local data samples, without the need to exchange the data itself [66]. This method allows models to learn from a broad dataset without compromising individual privacy. Importantly, federated learning facilitates the classification of skin conditions while maintaining stringent data security protocols. Rather than transmitting raw data to a central

server, only model updates are shared, significantly reducing the risk of data breaches.

Additionally, privacy-preserving AI techniques such as differential privacy and homomorphic encryption provide additional layers of data protection [67]. Differential privacy protects individuals by ensuring that the removal or addition of a single data point does not substantially impact the outcome of the analysis. Meanwhile, homomorphic encryption allows for secure data processing by allowing computations to be performed on encrypted data without the need for decryption, preserving the confidentiality and integrity of the data throughout the processing phase [67]. It is crucial to develop and implement more robust frameworks that not only meet current ethical standards but also anticipate future challenges in health data security and privacy.

Conclusion

This review extensively examines the integration of deep learning with dermoscopic imaging for melanoma detection, highlighting significant advancements and the potential to transform dermatological diagnostics. Despite these advancements, challenges remain in dataset diversity, ethical considerations, and standardized evaluation metrics. Enhanced accuracy in early melanoma detection through deep learning models and hybrid approaches promises improved patient outcomes but requires addressing biases, particularly in datasets that predominantly reflect fair-skinned populations. Ethical implementation in dermatological settings necessitates rigorous frameworks to ensure patient privacy, data security, and transparent AI decisions. Future research should focus on refining AI methodologies, ensuring seamless integration with clinical workflows, and more effectively incorporating clinician expertise into the decision-making process while maintaining patient autonomy. By addressing these areas, deep learning technologies can play a transformative role in dermatological diagnostics, particularly melanoma detection, offering more personalized and effective treatment options. This progression will not only enhance the diagnostic capabilities of dermatologists but also lead to better patient management and improved healthcare delivery.

References

1. Wang, J. Y., Wang, E. B., & Swetter, S. M. (2023). What Is Melanoma?. *JAMA*, *329*(11), 948. https://doi.org/10.1001/jama.2022.24888
2. Shaheen, F., Verma, B., & Asafuddoula, M. (2016). *Impact of Automatic Feature Extraction in Deep Learning Architecture*. https://doi.org/10.1109/DICTA.2016.7797053
3. Sonthalia, S., Yumeen, S., & Kaliyadan, F. (2023). Dermoscopy Overview and Extradiagnostic Applications. In *StatPearls*. StatPearls Publishing.
4. Errichetti E. (2020). Dermoscopy in Monitoring and Predicting Therapeutic Response in General Dermatology (Non-Tumoral Dermatoses): An Up-To-Date Overview. *Dermatology and therapy*, *10*(6), 1199–1214. https://doi.org/10.1007/s13555-020-00455-y
5. Duffy, K., & Grossman, D. (2012). The dysplastic nevus: from historical perspective to management in the modern era: part I. Historical, histologic, and clinical aspects. *Journal of the American Academy of Dermatology*, *67*(1), 1.e1–18. https://doi.org/10.1016/j.jaad.2012.02.047
6. Black, W. C., & Hunt, W. C. (1990). Histologic correlations with the clinical diagnosis of dysplastic nevus. *The American journal of surgical pathology*, *14*(1), 44–52. https://doi.org/10.1097/00000478-199001000-00005
7. Duarte, A. F., Sousa-Pinto, B., Azevedo, L. F., Barros, A. M., Puig, S., Malvehy, J., Haneke, E., & Correia, O. (2021). Clinical ABCDE rule for early melanoma detection. *European journal of dermatology: EJD*, *31*(6), 771–778. https://doi.org/10.1684/ejd.2021.4171
8. Elder D. E. (2006). Precursors to melanoma and their mimics: nevi of special sites. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, *19 Suppl 2*, S4–S20. https://doi.org/10.1038/modpathol.3800515
9. Dessinioti, C., Befon, A., & Stratigos, A. J. (2023). The Association of Nevus-Associated Melanoma with Common or Dysplastic Melanocytic Nevus: A Systematic Review and Meta-Analysis. *Cancers*, *15*(3), 856. https://doi.org/10.3390/cancers15030856
10. Friedman, R. J., Farber, M. J., Warycha, M. A., Papathasis, N., Miller, M. K., & Heilman, E. R. (2009). The "dysplastic" nevus. *Clinics in Dermatology*, *27*(1), 103–115. https://doi.org/10.1016/j.clindermatol.2008.09.008
11. Fried, L., Tan, A., Bajaj, S., Liebman, T. N., Polsky, D., & Stein, J. A. (2020). Technological advances for the detection of melanoma: Advances in diagnostic techniques. *Journal of the American Academy of Dermatology*, *83*(4), 983–992. https://doi.org/10.1016/j.jaad.2020.03.121
12. Srinidhi, C. L., Ciga, O., & Martel, A. L. (2021). Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, *67*, 101813. https://doi.org/10.1016/j.media.2020.101813
13. Luan, S., Chen, C., Zhang, B., Han, J., & Liu, J. (2018). Gabor Convolutional Networks. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 27(9), 4357–4366. https://doi.org/10.1109/TIP.2018.2835143
14. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
15. Canales-Fiscal, M. R., & Tamez-Peña, J. G. (2023). Hybrid morphological-convolutional neural networks for computer-aided diagnosis. Frontiers in artificial intelligence, 6, 1253183. https://doi.org/10.3389/frai.2023.1253183
16. Salehi, A., & Balasubramanian, M. (2023). DDCNet: Deep Dilated Convolutional Neural Network for Dense Prediction. Neurocomputing, 523, 116–129. https://doi.org/10.1016/j.neucom.2022.12.024
17. Sauter, D., Lodde, G., Nensa, F., Schadendorf, D., Livingstone, E., & Kukuk, M. (2023). Deep learning in computational dermatopathology of melanoma: A technical systematic literature review. *Computers in biology and medicine*, *163*, 107083. https://doi.org/10.1016/j.compbiomed.2023.107083
18. Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., & Chang, S. E. (2018). Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *The Journal of investigative dermatology*, *138*(7), 1529–1538. https://doi.org/10.1016/j.jid.2018.01.028
19. Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A., Braun, S. A., Carr, R.,

Combalia, M., Fernandez-Figueras, M. T., Ferrara, G., Fraitag, S., French, L. E., Gellrich, F. F., Ghoreschi, K., Goebeler, M., … Brinker, T. J. (2021). Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European journal of cancer (Oxford, England: 1990)*, *156*, 202–216. https://doi.org/10.1016/j.ejca.2021.06.049

20. Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A., Braun, S. A., Carr, R., Combalia, M., Fernandez-Figueras, M. T., Ferrara, G., Fraitag, S., French, L. E., Gellrich, F. F., Ghoreschi, K., Goebeler, M., … Brinker, T. J. (2021). Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European journal of cancer (Oxford, England: 1990)*, *156*, 202–216. https://doi.org/10.1016/j.ejca.2021.06.049
21. Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. BMC medical imaging, 22(1), 69. https://doi.org/10.1186/s12880-022-00793-7
22. Miñoza, J. M. A., Rico, J. A., Zamora, P. R. F., Bacolod, M., Laubenbacher, R., Dumancas, G. G., & de Castro, R. (2022). Biomarker Discovery for Meta-Classification of Melanoma Metastatic Progression Using Transfer Learning. Genes, 13(12), 2303. https://doi.org/10.3390/genes13122303
23. Annaby, M. H., Elwer, A. M., Rushdi, M. A., & Rasmy, M. E. M. (2021). Melanoma Detection Using Spatial and Spectral Analysis on Superpixel Graphs. Journal of digital imaging, 34(1), 162–181. https://doi.org/10.1007/s10278-020-00401-6
24. Schmid-Saugeon P. (2000). Symmetry axis computation for almost-symmetrical and asymmetrical objects: application to pigmented skin lesions. Medical image analysis, 4(3), 269–282. https://doi.org/10.1016/s1361-8415(00)00019-0
25. Qin, Z., Liu, Z., Zhu, P., & Xue, Y. (2020). A GAN-based image synthesis method for skin lesion classification. *Computer methods and programs in biomedicine*, *195*, 105568. https://doi.org/10.1016/j.cmpb.2020.105568
26. Guo, L., Xie, G., Xu, X., & Ren, J. (2020). Effective Melanoma Recognition Using Deep Convolutional Neural Network with Covariance Discriminant Loss. *Sensors (Basel, Switzerland)*, *20*(20), 5786. https://doi.org/10.3390/s20205786
27. Nagendram, S., Singh, A., Harish Babu, G., Joshi, R., Pande, S. D., Ahammad, S. K. H., Dhabliya, D., & Bisht, A. (2023). Stochastic gradient descent optimisation for convolutional neural network for medical image segmentation. *Open life sciences*, *18*(1), 20220665. https://doi.org/10.1515/biol-2022-0665
28. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, *5*, 180161. https://doi.org/10.1038/sdata.2018.161
29. Mohanty, N., Pradhan, M., Reddy, A. V. N., Kumar, S., & Alkhayyat, A. (2022). Integrated Design of Optimized Weighted Deep Feature Fusion Strategies for Skin Lesion Image Classification. *Cancers*, *14*(22), 5716. https://doi.org/10.3390/cancers14225716
30. Gao J. (2022). Network Intrusion Detection Method Combining CNN and BiLSTM in Cloud Computing Environment. *Computational intelligence and neuroscience*, *2022*, 7272479. https://doi.org/10.1155/2022/7272479
31. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747
32. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056
33. Vestergaard, M. E., Macaskill, P., Holt, P. E., & Menzies, S. W. (2008). Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, *159*(3), 669-676. https://doi.org/10.1111/j.1365-2133.2008.08713.x
34. Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J. Y., Kreusch, J., Lallas, A., Lapins, J., Marghoob, A., Menzies, S., Neuber, N. M., Paoli, J., Rabinovitz, H. S., Rinner, C., Scope, A., Soyer, H. P., Sinz, C., … Kittler, H. (2019). Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA dermatology*, *155*(1), 58–65. https://doi.org/10.1001/jamadermatol.2018.4378
35. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Reader study level-I and level-II Groups, Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-Billmann, T., Bowling, J., … Zalaudek, I. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology: official journal of the European Society for Medical Oncology*, *29*(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166
36. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, *4*(1), 65. https://doi.org/10.1038/s41746-021-00438-z
37. Khalifa, M., & Albadawy, M. (2024). AI in Diagnostic Imaging: Revolutionising Accuracy and Efficiency. *Computer Methods and Programs in Biomedicine Update*, *5*, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146
38. Jojoa Acosta, M. F., Caballero Tovar, L. Y., Garcia-Zapirain, M. B., & Percybrooks, W. S. (2021). Melanoma diagnosis using deep learning techniques on dermatoscopic images. *BMC medical imaging*, *21*(1), 6. https://doi.org/10.1186/s12880-020-00534-8
39. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence*, *31*. https://doi.org/10.1609/aaai.v31i1.11231
40. Singh, S.K., Banerjee, S., Chakraborty, A., Bandyopadhyay, A. (2023). Classification of Melanoma Skin Cancer Using Inception-ResNet. In: Mandal, J.K., De,

D. (eds) Frontiers of ICT in Healthcare. Lecture Notes in Networks and Systems, vol 519. Springer, Singapore. https://doi.org/10.1007/978-981-19-5191-6_6

41. Alwakid, G., Gouda, W., Humayun, M., & Jhanjhi, N. Z. (2023). Diagnosing Melanomas in Dermoscopy Images Using Deep Learning. *Diagnostics (Basel, Switzerland)*, *13*(10), 1815. https://doi.org/10.3390/diagnostics13101815
42. S M, J., P, M., Aravindan, C., & Appavu, R. (2023). Classification of skin cancer from dermoscopic images using deep neural network architectures. *Multimedia tools and applications*, *82*(10), 15763–15778. https://doi.org/10.1007/s11042-022-13847-3
43. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv*, *abs/1905.11946*.
44. Guo, L. N., Lee, M. S., Kassamali, B., Mita, C., & Nambudiri, V. E. (2022). Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *Journal of the American Academy of Dermatology*, *87*(1), 157–159. https://doi.org/10.1016/j.jaad.2021.06.884
45. Brady, J., Kashlan, R., Ruterbusch, J., Farshchian, M., & Moossavi, M. (2021). Racial Disparities in Patients with Melanoma: A Multivariate Survival Analysis. *Clinical, cosmetic and investigational dermatology*, *14*, 547–550. https://doi.org/10.2147/CCID.S311694
46. Mitre, M., Hosein, S., Mitri, A., Kurtansky, N. R., Mancebo, S. E., Fonseca, M., Jacobs, A. K., Rotemberg, V., & Marchetti, M. A. (2024). Dermatoscopic features and potential pitfalls of artificial intelligence-based analysis of benign acral pigmented lesions in Black patients: A multicenter observational study. *Journal of the American Academy of Dermatology*, S0190-9622(24)00502-4. Advance online publication. https://doi.org/10.1016/j.jaad.2024.02.058
47. Marchetti, M. A., Cowen, E. A., Kurtansky, N. R., Weber, J., Dauscher, M., DeFazio, J., Deng, L., Dusza, S. W., Haliasos, H., Halpern, A. C., Hosein, S., Nazir, Z. H., Marghoob, A. A., Quigley, E. A., Salvador, T., & Rotemberg, V. M. (2023). Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *NPJ digital medicine*, *6*(1), 127. https://doi.org/10.1038/s41746-023-00872-1
48. Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018, 4-7 April 2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018).
49. Wang, F., Kaushal, R., & Khullar, D. (2020). Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine?. *Annals of internal medicine*, *172*(1), 59–60. https://doi.org/10.7326/M19-2548
50. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6
51. Polesie, S., McKee, P. H., Gardner, J. M., Gillstedt, M., Siarov, J., Neittaanmäki, N., & Paoli, J. (2020). Attitudes Toward Artificial Intelligence Within Dermatopathology: An International Online Survey. *Frontiers in medicine*, *7*, 591952. https://doi.org/10.3389/fmed.2020.591952
52. Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H. P., Janda, M., Condon, J. J. J., Oakden-Rayner, L., Palmer, L. J., Keel, S., & van Wijngaarden, P. (2021). A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports*, *11*(1), 5193. https://doi.org/10.1038/s41598-021-84698-5
53. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, *abs/1312.6034*.
54. Sigut, J., Fumero, F., Estévez, J., Alayón, S., & Díaz-Alemán, T. (2023). In-Depth Evaluation of Saliency Maps for Interpreting Convolutional Neural Network Decisions in the Diagnosis of Glaucoma Based on Fundus Imaging. *Sensors (Basel, Switzerland)*, *24*(1), 239. https://doi.org/10.3390/s24010239
55. Dimanov, B. (2020). Interpretable Deep Learning: Beyond Feature-Importance with Concept-based Explanations [Apollo - University of Cambridge Repository]. https://doi.org/10.17863/CAM.73484
56. Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors (Basel, Switzerland)*, *23*(2), 634. https://doi.org/10.3390/s23020634
57. Kim, B., Wattenberg, M., Gilmer, J., Cai, C.J., Wexler, J., Viégas, F.B., & Sayres, R. (2017). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *International Conference on Machine Learning*.
58. Lin, W. C., Chen, J. S., Chiang, M. F., & Hribar, M. R. (2020). Applications of Artificial Intelligence to Electronic Health Record Data in Ophthalmology. *Translational vision science & technology*, *9*(2), 13. https://doi.org/10.1167/tvst.9.2.13
59. Andreotta, A. J., Kirkham, N., & Rizzi, M. (2022). AI, big data, and the future of consent. *AI & society*, *37*(4), 1715–1728. https://doi.org/10.1007/s00146-021-01262-5
60. Park H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. *Digital health*, *10*, 20552076241247938. https://doi.org/10.1177/20552076241247938
61. Adamson, A. S., & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA dermatology*, *154*(11), 1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348
62. Corbin, A., & Marques, O. (2023). Assessing Bias in Skin Lesion Classifiers with Contemporary Deep Learning and Post-Hoc Explainability Techniques. *IEEE Access*, *11*, 78339-78352. https://doi.org/10.1109/ACCESS.2023.3289320
63. Yap, J., Yolland, W., & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Experimental dermatology*, *27*(11), 1261–1267. https://doi.org/10.1111/exd.13777
64. Binder, M., Kittler, H., Dreiseitl, S., Ganster, H., Wolff, K., & Pehamberger, H. (2000). Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. *Melanoma research*, *10*(6), 556–561. https://doi.org/10.1097/00008390-200012000-00007

65. Alshahrani, M., Al-Jabbar, M., Senan, E. M., Ahmed, I. A., & Mohammed Saif, J. A. (2024). Analysis of dermoscopy images of multi-class for early detection of skin lesions by hybrid systems based on integrating features of CNN models. *PloS one*, *19*(3), e0298305. https://doi.org/10.1371/journal.pone.0298305
66. Riaz, S., Naeem, A., Malik, H., Naqvi, R. A., & Loh, W. K. (2023). Federated and Transfer Learning Methods for the Classification of Melanoma and Nonmelanoma Skin Cancers: A Prospective Study. *Sensors (Basel, Switzerland)*, *23*(20), 8457. https://doi.org/10.3390/s23208457
67. Sandhu, S. S., Gorji, H. T., Tavakolian, P., Tavakolian, K., & Akhbardeh, A. (2023). Medical Imaging Applications of Federated Learning. *Diagnostics (Basel, Switzerland)*, *13*(19), 3140. https://doi.org/10.3390/diagnostics13193140