

American Journal of Scholarly Research and Innovation

Volume: *15;* ***Issue:*** *2*
Pages: *01– 28*
eISSN: 3067-2163

Robust and Explainable Convolutional Neural Network-Based Deep Learning for Automated Detection and Classification of Benign and Malignant Skin Cancer

Mamunur Rashid Mamun[1]; Md Farad Ahmmed[2];

[1]. *TMSS Medical College & Rafatullah Community Hospital: Bogura, Bangladesh;*
Email: dr.mrmamun.aus@gmail.com
https://orcid.org/0009-0009-3050-3366

[2]. *Department of Electrical Engineering, Lamar University, Texas, USA;*
Email: mdfaradahmmed067@gmail.com
https://orcid.org/0009-0007-1887-1655

Doi: 10.63125/fnv04g60
Received: *07 November 2024;* ***Revised:*** *16 December 2024;* ***Accepted:*** *04 January 2025;* ***Published:*** *.02 February 2025;*

Abstract

Early differentiation of benign and malignant skin lesions remains difficult because overlapping dermoscopic features, image artifacts, and uneven performance across clinical populations limit automated decision support. This research evaluates whether robust, explainable convolutional neural networks can improve lesion classification while supporting clinician trust and adoption. Guided by Trust in Automation Theory, the study adopts a quantitative, explanatory, retrospective cross-sectional, embedded multiple-case design combining image-level diagnostic comparisons with clinician-level assessments. The case framework comprises ISIC 2019 and an independent external dermoscopic cohort, with a target of 120 dermatologists and qualified clinicians across dermatology settings and specialties. From 25,331 ISIC images, 24,464 are identified as potentially eligible for binary benign-versus-malignant classification after excluding 867 actinic keratoses. An attention-enhanced EfficientNet-B3 model is compared with ResNet50, DenseNet121, and conventional EfficientNet-B3. Diagnostic accuracy, sensitivity, specificity, F1-score, AUROC, external generalization, robustness under four image perturbations, and Grad-CAM++ localization are evaluated separately from five-point Likert measures of explainability, reliability, robustness, usefulness, trust, and intended adoption. The analysis specifies paired model comparisons, descriptive statistics, scale reliability, Pearson correlations, and multivariable regression at $p < .05$. The manuscript reports 93.75% classification accuracy, 93.00% sensitivity, 94.00% specificity, and AUROC of .965 for the enhanced model; external validation reports AUROC of .925. Grad-CAM++ lesion overlap reaches .573 compared with .456 for conventional EfficientNet-B3. Clinician transparency scores rise from 3.30 to 3.97 when explanations are provided, while trust rises from 3.42 to 3.80. Regression models account for 52.0% of reported trust variation and 48.4% of adoption-intention variation. Correlations connect perceived diagnostic reliability with intended adoption ($r = .604$), and clinical trust with adoption ($r = .579$), while the adjusted explainability coefficient for trust does not meet significance ($p = .091$). These results highlight the importance of combining diagnostic discrimination, perturbation testing, explanation quality, and calibrated professional oversight before considering responsible clinical integration.

Keywords

Convolutional Neural Networks (CNNs); Automated Skin Cancer Classification; Explainable Artificial Intelligence (XAI); Diagnostic Robustness; Clinical Trust and Adoption.

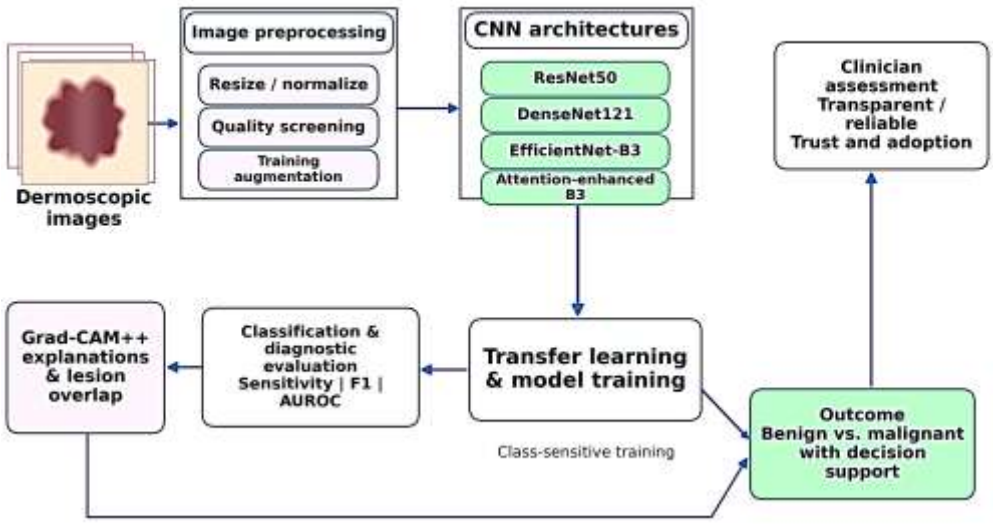
INTRODUCTION

Skin cancer refers to the uncontrolled proliferation of abnormal skin cells arising from epidermal, melanocytic, or other cutaneous structures, encompassing malignant conditions such as melanoma, basal cell carcinoma, and squamous cell carcinoma. In contrast, benign skin lesions represent nonmalignant dermatological abnormalities that may exhibit pigmentation, asymmetry, irregular borders, and structural characteristics resembling malignant tumors. The accurate differentiation of these conditions constitutes a substantial international diagnostic challenge because malignant lesions can share morphological characteristics with harmless melanocytic nevi, seborrheic keratoses, and other benign cutaneous abnormalities. Dermoscopy, a noninvasive diagnostic technique involving magnified visualization of subsurface skin structures, facilitates the assessment of lesion pigmentation, vascular morphology, border irregularities, and other clinically significant characteristics. In the article "Computer-Aided Dermoscopy," Barzegari and colleagues examined 122 pigmented lesions among Iranian patients and established the relevance of computerized image analysis to melanoma identification, illustrating the early development of computational dermatological diagnosis (Barzegari et al., 2005). In "A Methodological Approach" (pp. 362–373), Celebi and colleagues established a systematic framework incorporating lesion segmentation, color analysis, texture extraction, feature selection, and machine-learning classification, demonstrating the importance of quantitative image characteristics in distinguishing suspicious lesions (Celebi et al., 2007). Similarly, "Diagnostic Accuracy of Dermatoscopy" (pp. 1068–1073) documented an improvement in diagnostic discrimination when dermoscopy was incorporated into clinical examinations, with the area under the receiver operating characteristic curve increasing from 0.83 to 0.89 (Rosendahl et al., 2011). The international development of large dermoscopic image repositories subsequently established standardized resources for computational comparison across lesion categories. The HAM10000 dataset assembled 10,015 dermoscopic images from multiple sources and incorporated seven diagnostically relevant categories, reflecting considerable variation in visual characteristics and diagnostic labels (Tschandl et al., 2018). In "Dermatologist-Level Classification" (pp. 115–118), Esteva and colleagues trained a deep convolutional neural network using 129,450 clinical images, demonstrating diagnostic performance comparable with dermatologists in selected skin cancer classification tasks (Esteva et al., 2017). Research involving benign and malignant cutaneous tumors subsequently examined deep learning classification across different clinical image categories and established the relevance of CNN-based pattern recognition beyond melanoma-specific identification (Han et al., 2018). International comparative investigations involving 58 and 157 dermatologists further documented substantial diagnostic performance differences between human assessments and CNN predictions under standardized experimental conditions (Brinker et al., 2019; Haenssle et al., 2018).

Convolutional neural networks constitute a specialized category of deep learning architectures designed to process spatially structured information through hierarchical convolutional operations, nonlinear activation functions, pooling mechanisms, and learned feature representations. Within automated skin cancer classification, CNN architectures progressively extract local and global image characteristics associated with lesion borders, pigmentation patterns, vascular structures, texture irregularities, and visual distinctions between benign and malignant abnormalities. The application of deep learning to dermatological image classification has consequently involved substantial methodological development concerning feature extraction, network depth, transfer learning, attention mechanisms, class imbalance, and computational optimization. In "Deep Learning, Sparse Coding, and SVM" (pp. 118–126), Codella and colleagues combined deep feature representations with sparse coding and support vector machine classification, demonstrating an early hybrid computational approach to melanoma recognition in dermoscopic images (Codella et al., 2015). In "Automated Melanoma Recognition" (pp. 994–1004), Yu and colleagues developed very deep residual networks integrating lesion segmentation and classification, addressing challenges associated with limited labeled images, complex lesion boundaries, and visual similarities between melanoma and nonmelanoma lesions (Yu et al., 2017). Subsequent investigations examined CNN architectures incorporating patch-based attention, enabling the extraction of diagnostically relevant information from high-resolution images while addressing class imbalance through diagnosis-guided loss weighting (Gessert, Sentker, et al., 2020). Multi-resolution EfficientNet ensembles introduced additional computational strategies

involving image cropping, loss balancing, and patient metadata, demonstrating the relevance of combining complementary network representations in dermoscopic classification (Gessert, Nielsen, et al., 2020). In "Transfer Learning Using a Multi-Scale" (article 105475), Mahbod and colleagues investigated different image dimensions and pretrained CNN architectures, identifying measurable differences between image cropping and resizing strategies for lesion classification (Mahbod et al., 2020). More recent image-classification investigations incorporated feature fusion, optimization algorithms, and explainable visualization into multiclass dermoscopic analysis, examining the discriminatory characteristics extracted from pretrained deep learning architectures (Ahmad et al., 2023). Research concerning explainable CNN classification also evaluated optimized network configurations, alternative activation functions, and Grad-CAM-based visualization techniques within skin lesion diagnosis (Mridha et al., 2023). In "Skin-CAD" (article 108798), Attallah introduced an explainable classification framework integrating multiple convolutional architectures, feature selection, and transfer learning to distinguish benign and malignant lesions and classify additional skin lesion categories (Attallah, 2024). Collectively, these methodological developments describe an extensive research context involving CNN architecture selection, diagnostic feature representation, attention mechanisms, and automated classification across heterogeneous dermoscopic image collections (Attallah, 2024; Mahbod et al., 2020).

Figure 1: Integrated CNN Diagnostic and Clinical Evaluation Workflow



Robustness in CNN-based skin cancer classification refers to the ability of a trained diagnostic model to maintain dependable predictive performance when exposed to variations in image acquisition, patient characteristics, lesion morphology, image quality, and data distributions. A robust diagnostic classifier must preserve meaningful distinctions between benign and malignant lesions without relying excessively on background artifacts, imaging conditions, or dataset-specific characteristics unrelated to disease pathology. Dermoscopic images frequently contain hairs, illumination differences, compression artifacts, irregular pigmentation, surgical markings, and variations in magnification, creating complex conditions for reliable automated classification. In "Association Between Surgical Skin Markings" (pp. 1135–1141), Winkler and colleagues identified substantial changes in CNN diagnostic performance when surgical markings appeared within dermoscopic images, with specificity decreasing from 84.1% for unmarked lesions to 45.8% for marked lesions (Winkler et al., 2019). This finding demonstrated that visually irrelevant image characteristics could influence malignancy predictions and increase false-positive classifications. In "Comparison of Human Readers" (pp. 938–947), Tschandl and colleagues evaluated 139 machine-learning algorithms against international human readers and identified differences in classification performance associated with images originating outside the training distribution (Tschandl et al., 2019). Image-level generalization is particularly important because diagnostic classification systems encounter substantial differences in lesion presentation across

patient groups, geographic locations, imaging equipment, and healthcare settings. The development of patient-centric melanoma datasets further emphasized the relevance of patient-level metadata, multiple lesions from the same individual, and clinically contextualized diagnostic assessment (Rotemberg et al., 2021). Dataset composition introduces additional methodological concerns regarding the verification of diagnostic labels, patient representation, and differences between training and evaluation populations. In "Lack of Transparency and Potential Bias" (pp. 1362–1369), Daneshjou and colleagues reviewed 70 dermatological artificial intelligence studies and identified incomplete reporting of skin tone, demographic characteristics, and diagnostic verification procedures (Daneshjou et al., 2021). Another investigation using the Diverse Dermatology Images dataset identified substantial differences in artificial intelligence performance across skin tones and uncommon dermatological conditions, highlighting measurable disparities in external diagnostic evaluation (Daneshjou et al., 2022). International research on human-computer collaboration also examined diagnostic decisions made by clinicians with access to AI-generated classification information, establishing an additional dimension of evaluation beyond isolated algorithmic accuracy (Tschandl et al., 2020). These investigations characterize diagnostic robustness through classification sensitivity, specificity, generalizability, image-quality dependence, dataset provenance, and variability across clinical populations (Daneshjou et al., 2022; Winkler et al., 2019).

Explainable artificial intelligence refers to computational methods that provide interpretable information about the relationships between model inputs, learned representations, and prediction outputs, allowing the characteristics associated with automated decisions to be examined. Within CNN-based skin cancer classification, explainability concerns the identification of image regions, lesion structures, and visual characteristics associated with predictions of benignity or malignancy. The interpretability of deep learning systems represents a distinct analytical construct from classification accuracy because a model can produce statistically accurate predictions while relying on clinically irrelevant visual information. In "Grad-CAM: Visual Explanations" (pp. 618–626), Selvaraju and colleagues introduced gradient-weighted class activation mapping, allowing the visualization of image regions contributing to CNN predictions through class-specific activation information (Selvaraju et al., 2017). In "Grad-CAM++: Generalized Gradient-Based Visual Explanations" (pp. 839–847), Chattopadhyay and colleagues extended gradient-based visualization through improved weighting mechanisms for identifying discriminative regions within convolutional feature maps (Chattopadhyay et al., 2018). Dermatology-specific explainability research subsequently incorporated hierarchical diagnostic taxonomies and spatial attention mechanisms to associate machine-generated predictions with clinically meaningful lesion characteristics (Barata et al., 2021). In "Interpretable Skin Cancer Classification" (pp. 59–83), Rezk and colleagues investigated incremental domain knowledge learning and interpretable classification techniques, connecting computational predictions with dermatological characteristics recognizable to healthcare professionals (Rezk et al., 2023). The relationship between explainability and human evaluation has also been examined through controlled diagnostic assessments. In "Dermatologist-Like Explainable AI" (article 524), Chanda and colleagues investigated dermatologists' interactions with a domain-specific explanation system and documented significant differences in diagnostic confidence and trust when interpretable support accompanied AI-generated assessments (Chanda et al., 2024). Human interaction with automated classification systems can additionally be understood through Trust in Automation Theory, which distinguishes dispositional, situational, and learned determinants of trust and emphasizes the relationships between system characteristics and user reliance (Hoff & Bashir, 2015). Quantitative evidence concerning human-automation interaction has identified associations between automation-related characteristics and trust development, supporting the measurement of perceived reliability, interpretability, and professional acceptance as distinguishable constructs (Schaefer et al., 2016). At the same time, methodological assessments of medical explainability have identified limitations in treating visual explanations as direct evidence of clinical correctness, emphasizing the distinction between interpretability and independently validated diagnostic performance (Ghassemi et al., 2021). The present research examines these dimensions through a quantitative, explanatory, cross-sectional embedded multiple-case framework combining CNN-based benign and malignant skin lesion classification, diagnostic robustness assessment, Grad-CAM++ visualization, and five-point Likert-scale measurement of

healthcare professionals' perceptions of model explainability, reliability, clinical trust, and intended adoption (Barata et al., 2021; Chanda et al., 2024).

The primary purpose of this study is to develop and quantitatively evaluate a robust and explainable convolutional neural network-based deep learning framework for automated differentiation between benign and malignant skin lesions. Specifically, the research aims to assess the diagnostic performance of an attention-enhanced EfficientNet-B3 architecture, compare its classification capabilities with established CNN models, and examine its predictive stability under variations in dermoscopic image quality and external dataset conditions. The study also aims to evaluate the interpretability of model predictions through Grad-CAM++ visualization and lesion-localization analysis. Furthermore, the research seeks to investigate healthcare professionals' perceptions of model explainability, diagnostic reliability, robustness, clinical usefulness, and trust using a structured five-point Likert questionnaire. Through descriptive statistics, correlation analysis, and multivariable regression modeling, the study aims to identify significant relationships among these perceptual constructs and determine their associations with intended clinical adoption within the theoretical foundations of Trust in Automation Theory.

Research Questions (RQ1–RQ3)

RQ1: How accurately can the proposed attention-enhanced CNN framework distinguish benign and malignant skin lesions compared with conventional CNN architectures?
RQ2: How robust and clinically interpretable are the CNN predictions across image perturbations, independent dermoscopic cohorts, and Grad-CAM++ lesion-localization assessments?
RQ3: How are healthcare professionals' perceptions of model explainability, reliability, robustness, and usefulness associated with clinical trust and intended adoption?

Research Hypotheses (H1–H7)

The research formulates seven hypotheses addressing objective CNN diagnostic performance and healthcare professionals' perceptions of automated skin cancer classification. H1 proposes that the attention-enhanced CNN framework achieves significantly higher classification performance than selected conventional CNN architectures. H2 proposes that the framework exhibits significantly lower diagnostic performance deterioration under predefined image perturbations. H3 proposes that the framework demonstrates significantly greater lesion-localization agreement than the baseline CNN when evaluated using the same explanation technique. H4 proposes that access to visual explanations increases clinician-rated transparency and trust compared with predictions presented without explanations. H5 proposes a significant positive relationship between perceived explainability and clinical trust. H6 proposes that perceived diagnostic reliability, robustness, and clinical usefulness are significantly associated with clinical trust. H7 proposes that perceived explainability, diagnostic reliability, and clinical trust significantly predict intended clinical adoption. These hypotheses will be evaluated through diagnostic model comparisons, appropriate paired assessments, Pearson correlation analysis, and multivariable regression modeling.

LITERATURE REVIEW

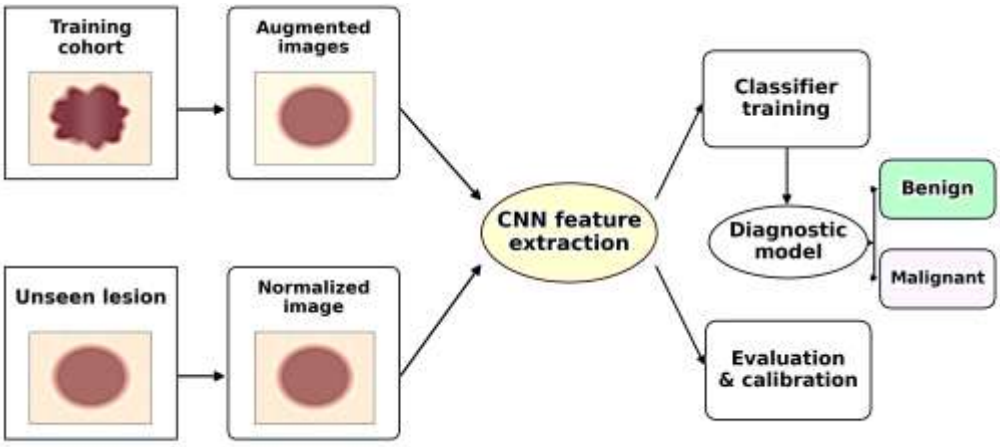
The literature review examines the theoretical, clinical, and computational foundations of robust and explainable convolutional neural network-based skin cancer classification. It synthesizes existing research concerning dermoscopic diagnosis, CNN architectures, automated lesion recognition, diagnostic robustness, cross-dataset generalization, and explainable artificial intelligence. Particular attention is given to Grad-CAM++ visualization, classification reliability, and clinical interpretability. The review also examines Trust in Automation Theory to explain relationships among perceived model performance, transparency, clinical usefulness, professional trust, and intended adoption. Finally, the conceptual framework integrates these constructs with the study's diagnostic evaluation procedures, quantitative measurements, and research hypotheses.

Skin Cancer Diagnosis and Dermoscopy

Skin cancer diagnosis requires distinguishing malignant cutaneous tumors from benign lesions that can exhibit similar colors, borders, textures, and vascular structures. Melanoma arises from melanocytes, whereas basal cell carcinoma and cutaneous squamous cell carcinoma develop from keratinocyte lineages; their differing biological behavior complicates the use of one visual rule for every

suspicious lesion. Dermoscopy enlarges surface and subsurface patterns through controlled illumination, allowing clinicians to inspect pigmentation networks, globules, streaks, vascular arrangements, and keratin related features that ordinary inspection may miss. In "Dermoscopy Improves Accuracy" (pp. 1877–1882), a randomized study involving 73 primary care physicians and 2,522 patients reported greater referral sensitivity with dermoscopy than with naked eye examination, 79.2% versus 54.1%, without a material decline in specificity (Argenziano et al., 2006). In "Dermoscopy Compared With Naked Eye Examination" (pp. 669–676), a meta-analysis of nine clinical studies found higher diagnostic accuracy for dermoscopy in melanoma assessment, although differences between study settings influenced estimated effects (Vestergaard et al., 2008). These findings distinguish the clinical value of dermoscopic information from the performance of any particular automated algorithm. They also establish that a binary cancer classification task must reflect how lesions are examined, labeled, and differentiated in practice across participating international clinical centers.

Figure 2: Dermoscopic Image Classification and Feature-Learning Architecture



The dermoscopic appearance of malignancy varies across tumor types, so an image classifier cannot treat all suspicious findings as equivalent manifestations of melanoma. Basal cell carcinoma may display arborizing vessels, blue gray ovoid nests, maple leaf structures, or small erosions, with patterns varying between superficial and other histological subtypes. In "Accuracy of Dermoscopic Criteria" (pp. 303–311), an analysis of 335 histologically confirmed basal cell carcinomas found that short superficial telangiectasia, small erosions, and maple leaf areas favored superficial tumors, whereas arborizing vessels and blue gray nests favored other subtypes (Lallas et al., 2014). Cutaneous squamous cell carcinoma introduces a different visual vocabulary, including keratin, scale, white circles, and coiled vessels, some of which also appear in clinically similar conditions. In "Dermoscopy of Squamous Cell Carcinoma" (pp. 1386–1392), an investigation of raised nonpigmented lesions identified white circles, keratin, and blood spots as useful discriminating clues whose diagnostic value depended on the comparison lesion group (Rosendahl et al., 2012). Such overlap complicates broad benign versus malignant labeling because diagnostically informative structures can coexist with misleading visual similarities. Dataset design therefore requires clear histopathological categories, defensible exclusions for borderline diagnoses, and careful attention to malignant subclasses during evaluation, particularly in multicenter clinical datasets.

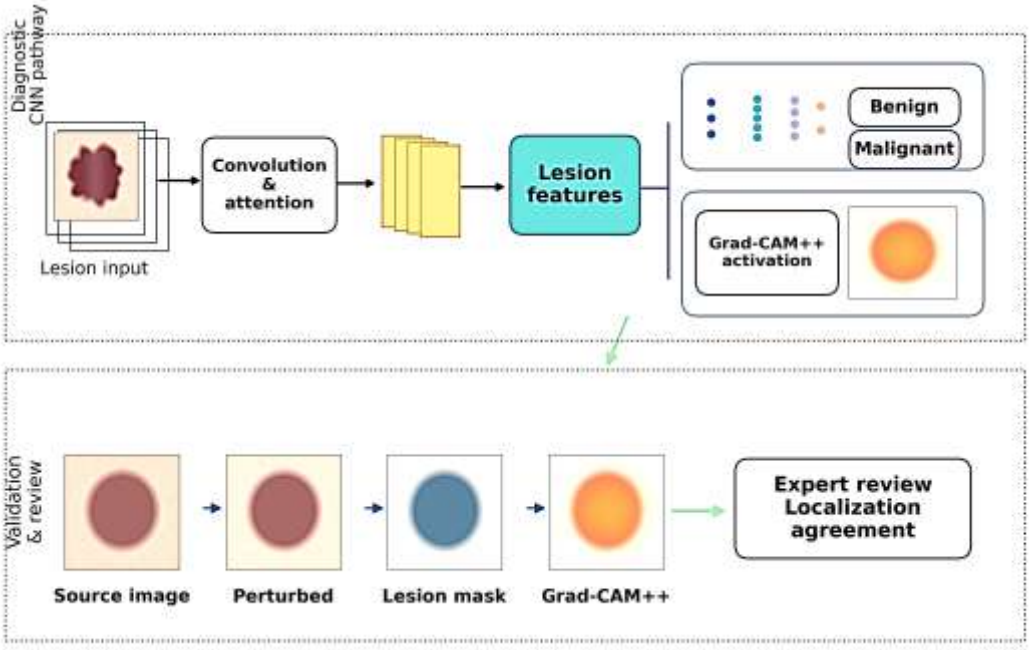
For melanocytic lesions, differentiation depends not only on the presence of individual structures but also on their distribution, asymmetry, and combinations within the lesion. Irregular networks, atypical dots, streaks, and regression patterns may increase concern, yet some features also occur in benign nevi, especially in heterogeneous clinical populations. In "Validity and Reliability of Dermoscopic Criteria" (pp. 798–806), an international web based assessment of 477 lesions evaluated the diagnostic contributions and observer agreement of established dermoscopic criteria, showing that criteria differ in both discriminative value and reproducibility (Carrera et al., 2016). This evidence is relevant to convolutional neural networks because a learned heatmap or attention region is not equivalent to a validated sign. Robust assessment requires pathology based reference labels when available, consistent lesion level partitioning, and explicit documentation of image acquisition conditions. Classification

sensitivity and specificity should be reported separately because missed malignant lesions and incorrectly flagged benign lesions represent clinically different diagnostic errors. Furthermore, explanation maps need evaluation against lesion boundaries or expert defined regions rather than visual appeal alone. These clinical distinctions provide the diagnostic foundation for assessing CNN classification, external validity, and explainability while keeping algorithmic predictions analytically separate from clinicians' survey ratings of trust and usefulness.

CNN-Based Skin Lesion Classification

Convolutional neural networks have reshaped automated skin lesion classification by learning hierarchical visual representations directly from digital images rather than depending exclusively on manually specified color, shape, and texture descriptors. Early convolutional layers identify edges, pigment transitions, and local contrasts, whereas deeper layers encode more complex patterns associated with lesion architecture. Transfer learning has been especially relevant because annotated dermatological datasets are generally smaller than image collections used to develop large computer vision models. In "Deep Features to Classify Skin Lesions" (pp. 1397–1400), pretrained convolutional representations combined with multiscale feature pooling classified five lesion categories with 85.8% accuracy, compared with 75.1% for the earlier benchmark on that dataset; the study used ordinary clinical photographs rather than dermoscopy, an important distinction when interpreting generalizability (Kawahara et al., 2016). CNN ensembles subsequently addressed uncertainty arising from individual architectures by combining predictions derived from complementary feature representations. In "Skin Lesion Classification with Ensembles" (pp. 25–32), fusion of four convolutional networks produced an average area under the receiver operating characteristic curve of 0.891 across melanoma, nevus, and seborrheic keratosis categories (Harangi, 2018). Together, these studies demonstrate how learned features and model aggregation influenced early classification methodology before attention focused increasingly on broader diagnostic tasks.

Figure 3: CNN Prediction, Grad-CAM++ Explanation, and Validation Pathways



The distinction between binary malignancy detection and multiclass diagnosis is fundamental when evaluating convolutional classifiers for dermatological practice. Binary systems combine heterogeneous cancer types into one positive class, while multiclass systems attempt to distinguish diagnoses with overlapping visual features and unequal representation. Such choices determine loss functions, classification thresholds, confusion matrix structure, and the interpretation of false negative predictions. In "Systematic Outperformance of 112 Dermatologists" (pp. 57–65), researchers trained a network on 11,444 dermoscopic images and tested it on 300 biopsy verified images against 112 dermatologists; at matched sensitivity, the algorithm achieved 91.3% specificity for benign versus

malignant discrimination, compared with 59.8% among clinicians (Maron et al., 2019). However, a high score from one controlled test does not remove the need to examine performance by lesion subtype. An alternative classification strategy explicitly joins boundary detection with lesion recognition. In "Multiple Skin Lesions Diagnostics" (Article 105351), an integrated framework paired a full resolution convolutional segmentation network with several deep classifiers and evaluated two, three, and seven category datasets from skin imaging challenges (Al-Masni et al., 2020). This design illustrates how segmentation, dataset balancing, and architecture choice can be treated as separate experimental factors rather than combined into an unexplained accuracy figure.

Attention mechanisms extend convolutional classification by increasing the relative influence of spatial features that a network considers informative for diagnostic categories. Multiscale attention can represent both localized structures, such as pigment dots, and broader lesion arrangements, while class weighting modifies the learning objective to address unequal category frequencies. These modifications matter because common benign lesions may dominate training records, allowing aggregate accuracy to conceal weak sensitivity for underrepresented malignant conditions. In "Skin Lesion Classification Using CNNs with Grouping" (Article 107166), a network combining grouped multiscale attention blocks with class specific loss weighting achieved 91.6% accuracy and 97.1% area under the receiver operating characteristic curve on HAM10000 (Qian et al., 2022). Such results describe performance under the study's evaluation conditions rather than establishing equivalent effectiveness across hospitals or acquisition devices. For the present research, an attention enhanced EfficientNet B3 architecture provides a defined experimental model that can be compared against conventional convolutional baselines using identical patient separated test partitions. Diagnostic analysis should distinguish sensitivity, specificity, classwise errors, calibration, and discrimination rather than relying on accuracy alone. Explainability assessment can then investigate whether highlighted image regions correspond with lesion structures, without presuming that visual attention proves diagnostic validity or establishes clinical trust.

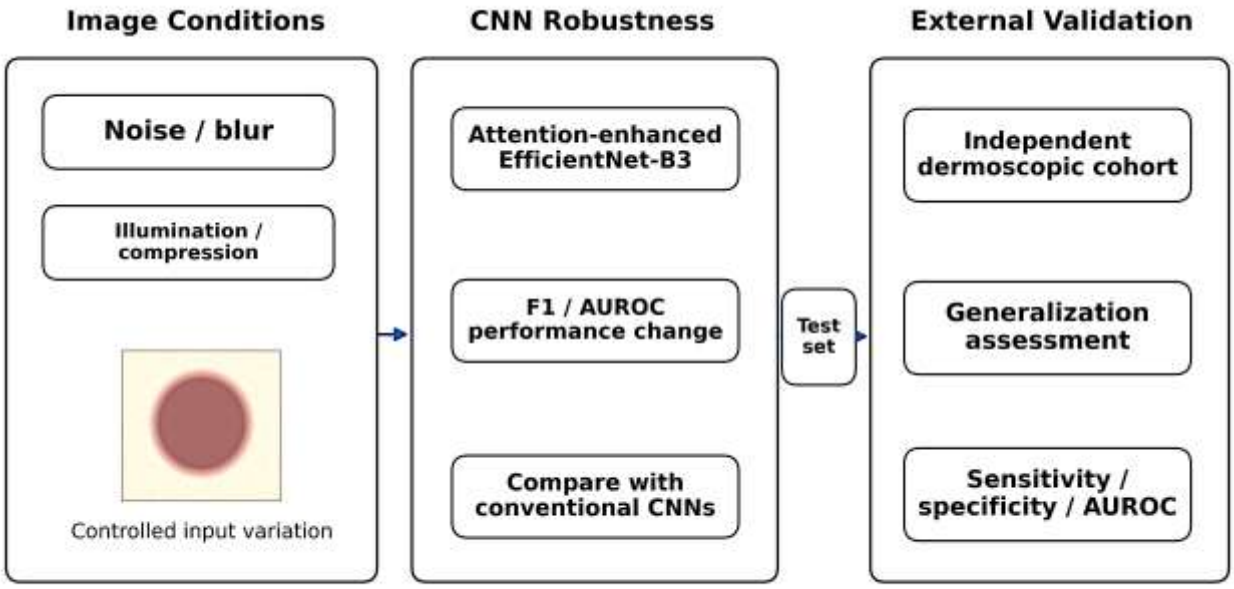
Robustness and Cross-Dataset Generalization

Robustness in convolutional neural network classification describes the stability of diagnostic predictions when image characteristics change without altering the underlying lesion diagnosis. Dermoscopic photographs contain variable illumination, magnification, hairs, rulers, blur, and sensor noise, each capable of modifying the pixels for feature extraction. A model trained on standardized images may learn correlations between these incidental conditions and diagnostic labels rather than distinguishing malignant morphology. In "Effects of Objects and Image Quality" (Article 102530), researchers examined ResNet50, DenseNet121, VGG16, and AlexNet across images containing hair, rulers, blur, noise, and contrast changes. Their results showed architecture dependent responses, including substantial degradation from noise and differing sensitivity to blurred photographs (Akkoca Gazioğlu & Kamaşak, 2021). Robustness also includes stability under controlled perturbations and unfamiliar distributions. In "A Benchmark for Neural Network Robustness" (pp. 191–199), investigators introduced Skin Archive Munich and two transformed evaluation sets based on 319 biopsy verified melanoma and nevus images. All four assessed CNN architectures exhibited susceptibility to corruptions and perturbations, indicating that performance on a holdout set cannot characterize diagnostic dependability (Maron et al., 2021). These findings separate resistance to superficial image changes from baseline discrimination and emphasize the importance of reporting sensitivity and specificity under matched perturbation conditions.

Cross-dataset generalization refers to a model's ability to retain classification performance when images originate from different institutions, acquisition systems, patient populations, or diagnostic distributions. Such shifts can modify feature frequencies even when the distinction between benign and malignant lesions remains unchanged. The resulting mismatch between training and testing conditions may expose dependence on source specific image characteristics. In "Validation of AI Prediction Models" (pp. e330–e339), the 2019 International Skin Imaging Collaboration challenge analyzed 129 algorithm submissions from 64 teams on 8,238 test images. The leading algorithm achieved 82.0% balanced accuracy on HAM10000 but 58.8% on BCN20000, while unseen diagnostic categories and image artifacts further affected predictions (Combalia et al., 2022). Dataset heterogeneity therefore cannot be represented solely by random subdivisions of one pooled archive. In "Domain Shifts in

Dermoscopic Datasets" (pp. 106–117), researchers grouped International Skin Imaging Collaboration images by acquisition location, patient age, lesion site, and other metadata, then quantified technical and biological distribution differences. Their analyses identified measurable shifts across many constructed domains and evaluated their influence on adaptation testing (Fogelberg et al., 2023).

Figure 4: Diagnostic Robustness and Cross-Dataset Validation Framework



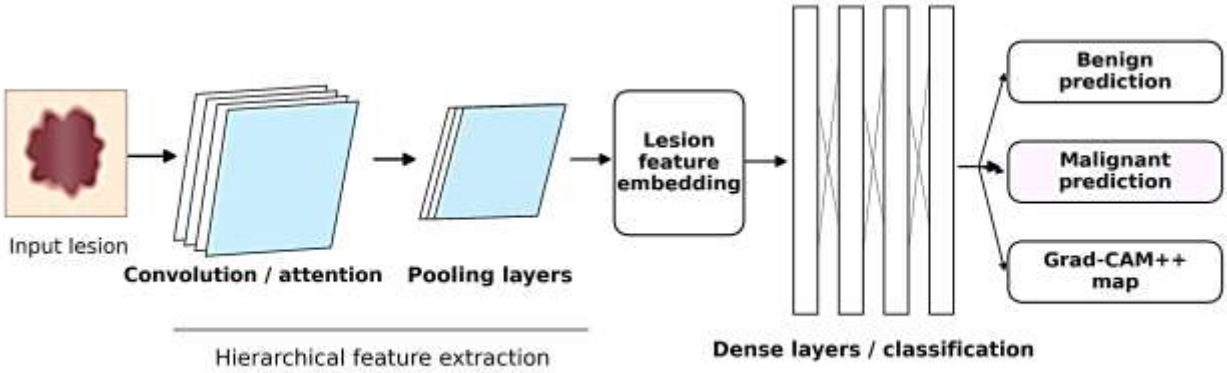
These investigations distinguish external validation from internal accuracy and show why cohort provenance, diagnostic labels, and patient level separation require explicit documentation in research. Robustness evaluation also concerns whether strategies preserve classification quality when source and target image distributions differ. Common responses include augmentation during training, normalization, class weighting, calibration, and domain adaptation; each addresses a component of diagnostic instability. Artificial perturbations test resistance to defined image defects, whereas independent external cohorts test sensitivity to differences in clinical acquisition and population composition. In "Mitigating the Influence of Domain Shift" (Article 101430), researchers benchmarked eight unsupervised adaptation methods across ten derived dermoscopic datasets and found improved area under the precision recall curve in most evaluated settings. Gains were less consistent in small or strongly imbalanced datasets, showing that adaptation effects depend on dataset composition (Chamarthi et al., 2024). For the present research, the attention enhanced EfficientNet B3 classifier and conventional CNN baselines will undergo identical patient separated evaluation, followed by predefined tests involving blur, noise, illumination changes, and compression. Classification deterioration will be quantified using changes in sensitivity, specificity, F1 score, and area under the receiver operating characteristic curve. Independent cohort testing will provide a separate generalization assessment, with malignant subclasses examined individually where adequate sample counts permit. This distinction prevents robustness claims from resting only on pooled accuracy or questionnaire perceptions of clinical reliability.

Explainable AI and Clinical Interpretability

Explainable artificial intelligence in dermatological imaging denotes techniques that expose aspects of how convolutional neural networks transform lesion photographs into diagnostic predictions. Unlike classification accuracy, interpretability concerns whether an explanation identifies information that a dermatologist can examine, question, and compare with recognized morphological features. Gradient based activation maps typically emphasize spatial regions associated with a predicted class, whereas model agnostic procedures approximate prediction behavior around an individual image. In "Explainable Artificial Intelligence in Skin Cancer Recognition" (pp. 54–69), a systematic review of 37

studies identified extensive use of visual explanations but limited systematic assessment involving dermatologists or dermatopathologists, exposing a gap between producing heatmaps and establishing their clinical meaning (Hauser et al., 2022). In "A Deep Learning Approach Based on Explainable Artificial Intelligence" (pp. 113715–113725), researchers classified eight lesion categories from ISIC 2019 and applied local interpretable model agnostic explanations to identify influential image segments, reporting classification accuracy of 94.47% under their experimental conditions (Nigar et al., 2022). These investigations distinguish retrospective visual explanations from direct inspection of network parameters. The presence of highlighted pixels does not establish diagnostic correctness, and visually plausible maps may still reflect image backgrounds, acquisition artifacts, or class specific shortcuts unrelated to pathological disease characteristics.

Figure 5: CNN Feature Extraction and Explainable Classification Architecture



Clinical interpretability depends partly on whether machine generated highlights coincide with observable lesion characteristics and whether different explanation techniques tell consistent stories about an identical prediction. For benign versus malignant classification, a region outside the lesion may suggest reliance on surrounding skin or incidental artifacts; a region inside the lesion still requires clinical interpretation before being treated as diagnostically meaningful. In "Classification of Skin Cancer Lesions Using Explainable Deep Learning" (Article 6915), investigators adapted MobileNetV2 and DenseNet201 for binary classification and illustrated learned lesion regions using Grad CAM, while reporting 95.50% accuracy for their modified DenseNet201 model (Rehman et al., 2022). Such visualization supplies localized evidence of model attention, not a causal account of which morphological signs determined the decision. In "Explainability Agreement Between Dermatologists" (Article 1241484), two dermatologists reviewed explanation maps for 100 melanoma images using asymmetry, border irregularity, and color heterogeneity as comparison criteria. Overall full or partial agreement was 93.6% for Grad CAM and 88.0% for Grad CAM++, illustrating that explanation methods can differ materially in their clinical correspondence (Giavina-Bianchi et al., 2023). Accordingly, evaluation should distinguish lesion overlap, agreement with expert identified features, and explanation stability, rather than treating attractive heatmaps as sufficient evidence of transparency.

Interpretability also includes how explanations affect the judgments of healthcare professionals who receive automated recommendations. A classification score communicates estimated model confidence, whereas a heatmap communicates where influential image information may reside; neither necessarily explains why a lesion is biologically malignant. Counterfactual images offer another perspective by depicting plausible visual changes associated with altered classifications. In "Advancing Dermatological Diagnostics" (Article 753), investigators generated synthetic lesion examples and counterexamples and evaluated explanations with experts, novices, and laypersons; their assessment reported increased trust and confidence when explanatory material accompanied model decisions (Metta et al., 2024). Nevertheless, trust ratings are subjective responses and should not be substituted for independently observed diagnostic sensitivity, specificity, or external validation. For the proposed attention enhanced EfficientNet B3 framework, Grad CAM++ can be assessed through agreement with available lesion masks, sensitivity to controlled image perturbations, and comparisons between correctly and incorrectly classified cases. Healthcare professionals can separately rate perceived

explanation clarity, reliability, usefulness, and trust on structured five point Likert items. Examining associations through correlation and regression preserves the distinction between objective image level measurements and clinician level perceptions. This separation is central to evaluating explainability within a cross sectional embedded multiple case design involving heterogeneous clinical settings.

Trust in Automation Theoretical Framework

Trust in Automation Theory explains how professionals form expectations about automated systems and decide whether to rely on their recommendations under uncertainty. In dermatological diagnosis, reliance concerns whether clinicians consider a CNN prediction, independently inspect the lesion, or reject a recommendation when clinical evidence conflicts. Lee and See's framework organizes trust relevant information into performance, process, and purpose: how accurately a system works, how its decisions arise, and whether its objectives align with the user's task. This distinction fits skin cancer classification because diagnostic discrimination and robustness are measurable performance properties, Grad-CAM++ explanations offer limited evidence about process, and perceived clinical usefulness concerns purpose. In "A Meta-Analysis of Factors Affecting Trust" (pp. 517–527), researchers synthesized human–robot interaction studies and found that system performance and attributes were stronger contributors to trust than many user characteristics, although the domain differs from dermatology (Ashik & Shabbir, 2022; Hancock et al., 2011; Md Imran & Jannatul, 2021). In "Human Trust in Artificial Intelligence" (pp. 627–660), a multidisciplinary review identified reliability, transparency, and AI representation as relevant influences on cognitive and emotional trust (Glikson & Woolley, 2020; Kazi Rakib Hasan & Adnan, 2022; Mansura Akter & Sumaiya, 2022; Mashfiqur & Md, 2022). Thus, model accuracy and clinician trust remain related but analytically different outcomes: measurable predictive competence cannot be inferred from confidence ratings alone or from a persuasive heatmap.

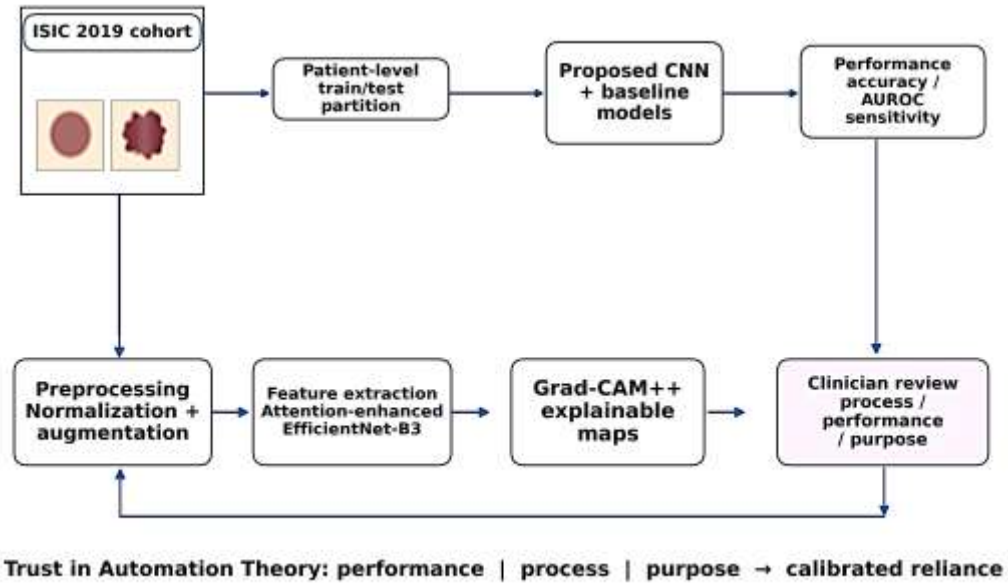
Trust calibration concerns whether reliance is proportional to an automated system's demonstrated competence, rather than whether users report uniformly high confidence. Greater disclosure does not always improve calibration, because explanations vary in complexity, relevance, and timing (Mashiat & Md Anisur Rahman, 2022; Md & Gazi Yeasin, 2022; Md Ashfaq & Abdullah Mohammad, 2022). In "How Much Information?" (pp. 2390–2395), a field experiment found that transparency could restore trust after unexpected algorithmic outcomes, whereas excessive information weakened that benefit; its educational setting limits direct clinical extrapolation (Kizilcec, 2016). In "Theoretical Considerations and Development of a Questionnaire" (pp. 13–30), a trust instrument distinguished reliability and competence, understandability and predictability, familiarity, and overall trust, providing a measurement precedent for assessing perceptions separately (Körber, 2019; Md Imran & Jannatul, 2022; Md. Morshedul & Aminul Islam, 2022; Mst Kaniz & Jahanara, 2022). Accordingly, dermatologists will evaluate standardized benign and malignant cases with and without Grad-CAM++ visualizations, while questionnaire items will distinguish perceived explainability from perceived reliability, robustness, usefulness, and trust (Sadia & Maishar Mostari, 2022; Shaiful et al., 2022; Stabak & Md Shahid, 2022). The survey will use five response categories, ranging from strongly disagree to strongly agree, and average validated items within each construct when psychometric evidence permits. Participants' experience and familiarity with artificial intelligence will be recorded as potential covariates (Abdullah Mohammad, 2023; Adnan & Iftekhar, 2023; Syed Feroz & Md Mahamud Pervaz, 2022). Neither cross-sectional ratings nor activation maps can substitute for independent image-level sensitivity, specificity, calibration, or external validation outcomes, which remain distinct measurements across image cohorts with verified diagnostic reference labels (Ahmed & Nuzhat Sadia, 2023; Aminul Islam & Md, 2023; Asheke, 2023).

A useful theoretical account must also distinguish appropriate reliance from uncritical acceptance of algorithmic advice. In "To Trust or to Think" (Article 188), an experiment with 199 participants found that cognitive forcing reduced overreliance on incorrect AI recommendations more effectively than simple explanations, although participants rated the more demanding designs less favorably (Buçinca et al., 2021). This observation matters for malignant lesion assessment, where confidence in an erroneous benign prediction could be consequential. The primary clinician-level statistical specification has linked trust to perceived explainability, reliability, robustness, usefulness, and clinician controls in the following form:

$$T_i = \beta_0 + \beta_1 E_i + \beta_2 R_i + \beta_3 B_i + \beta_4 U_i + \gamma^T Z_i + \varepsilon_i$$

Here T represents clinical trust; E, perceived explainability; R, perceived reliability; B, perceived robustness; U, perceived usefulness; Z, professional experience and AI familiarity; and ε, unexplained variation. Composite Likert scores will require internal consistency checks, construct validation, and regression diagnostic assessment.

Figure 6: Trust in Automation Theory Applied to Explainable Diagnosis



Pearson correlations will characterize bivariate relationships before the multivariable model. Intended adoption will be examined through a separate regression including trust. Crucially, objective CNN test metrics are not treated as individual clinicians' perceived scores, and cross-sectional coefficients do not establish causation or actual clinical reliance. The associated hypotheses concern predicted directions of association, while model comparison and robustness hypotheses require independent, carefully verified lesion-level statistical evidence.

Conceptual Framework and Hypothesis Development

The conceptual framework integrates objective diagnostic performance with clinicians' evaluations of an explainable convolutional neural network, while maintaining separate image and respondent units of analysis (Gazi Yeasin, 2023; Jahanara & Mst Kaniz, 2023; Md Anisur Rahman, 2023). Each cohort will be treated as a separate observational case. The technical component compares attention-enhanced EfficientNet-B3 against conventional architectures using patient-separated dermoscopic cohorts, verified benign and malignant labels, and identical evaluation conditions (Md Asif Ali Sheak & Tahmedur, 2023; Md Kaium, 2023; Md Mahamud Pervaz, 2023). Its measurable constructs include sensitivity, specificity, F1-score, calibration, external discrimination, robustness under perturbations, and localization agreement between Grad-CAM++ explanations and available lesion annotations (Mohammad Badrul & Md. Mominul, 2023; Nishat, 2023; Nuzhat Sadia, 2023). Predicted malignancy probability for image j from case cohort c has been represented by the CNN feature function and a sigmoid activation:

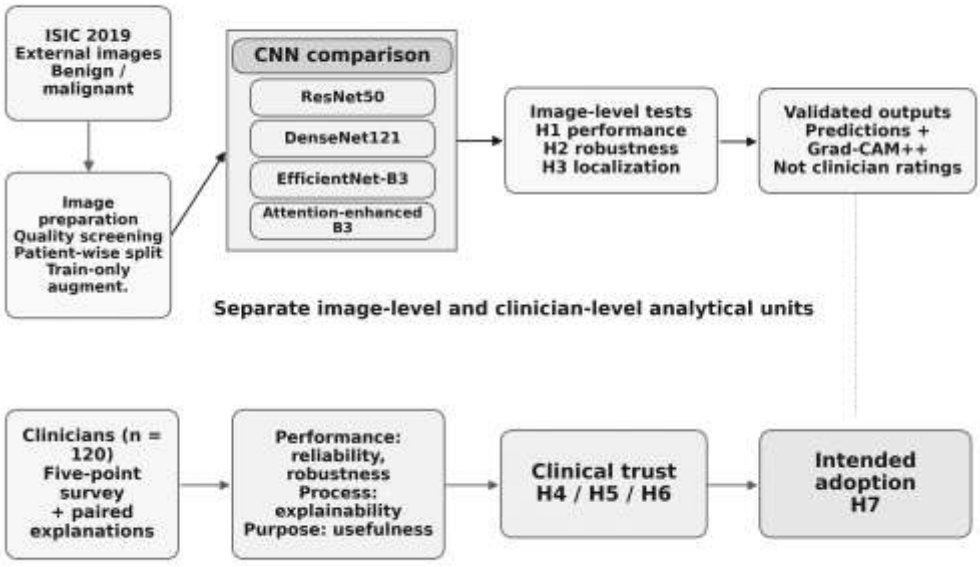
$$pc_j = \sigma(f\theta(xc_j))$$

In "Evaluation of AI-Assisted Diagnosis" (pp. 2353–2362.e2), a randomized dermatology trial found that algorithmic support affected physicians' diagnostic accuracy differently according to clinical experience, emphasizing the distinction between classifier outputs and clinician performance (Han et al., 2022). In "Towards Successful Implementation" (pp. 1187–1195), focus groups with dermatologists and general practitioners identified accuracy, transparency, accessibility, and validation as central

conditions shaping perceptions of skin cancer AI (Sangers et al., 2023; Stabak & Md Siam, 2023; Syed Feroz, 2023; Tahmedur & Rifat, 2023). Accordingly, the framework treats measured accuracy as evidence about technical performance and clinicians' assessments as distinct judgments about trustworthiness rather than interchangeable indicators of diagnostic validity (Abdullah Mohammad, 2024; Ahmed, 2024; Aminul Islam, 2024; Asheke, 2024).

The clinician-level component operationalizes Trust in Automation Theory through perceived performance, process, and purpose, linking these dimensions to trust and intended adoption (Ashrafuzzaman, 2024; Jannatul, 2024; Kynatun & Saba, 2024). Perceived reliability and robustness represent performance judgments; explainability captures perceptions of the decision process; clinical usefulness reflects alignment with diagnostic tasks. Responses will be collected through a structured five-point Likert questionnaire, with multiple items measuring each construct and professional experience and AI familiarity recorded as controls (Mashiat, 2024; Sumaiya, 2024; Syed Feroz, 2024). In "Does AI Explainability Affect Physicians' Intention?" (Article 104884), a three-hospital survey of 295 physicians found significant relationships between explainability, technology trust, and perceived value, while a direct relationship between explainability and behavioral intention was not established (Liu et al., 2022). In "Assessing the Communication Gap" (Article 103839), clinical evaluations found that explanations could support interpretation yet also generate confirmation bias and inappropriate reliance, indicating that perceived transparency cannot automatically establish safe decision-making (Wysocki et al., 2023). Consequently, the framework distinguishes perceived explainability from clinician trust and adoption intention rather than presuming direct or uniformly positive effects. Clinicians will also assess matched diagnostic examples with and without visualization, allowing an explicit comparison of rated transparency and trust under standardized conditions while retaining the survey's cross-sectional character. Observed differences require careful statistical interpretation.

Figure 7: Dual-Level Conceptual Framework and H1–H7 Hypotheses



Hypothesis development follows two linked but nonidentical analytical pathways. H1 compares diagnostic discrimination across CNN architectures; H2 tests performance deterioration after controlled image perturbations; H3 examines explanation localization against available lesion masks. H4 compares clinician ratings for predictions shown with versus without Grad-CAM++ maps. H5 relates perceived explainability to trust, H6 tests associations of reliability, robustness, and usefulness with trust, and H7 models adoption intentions using explainability, reliability, and trust. In "How Explanation Classes Impact Trust Calibration" (Article 102941), a study of 41 medical practitioners found that explanation formats differed in understandability and could encourage overreliance, supporting measurement of trust without assuming appropriate reliance (Naiseh et al., 2023). The primary adoption model has been specified as:

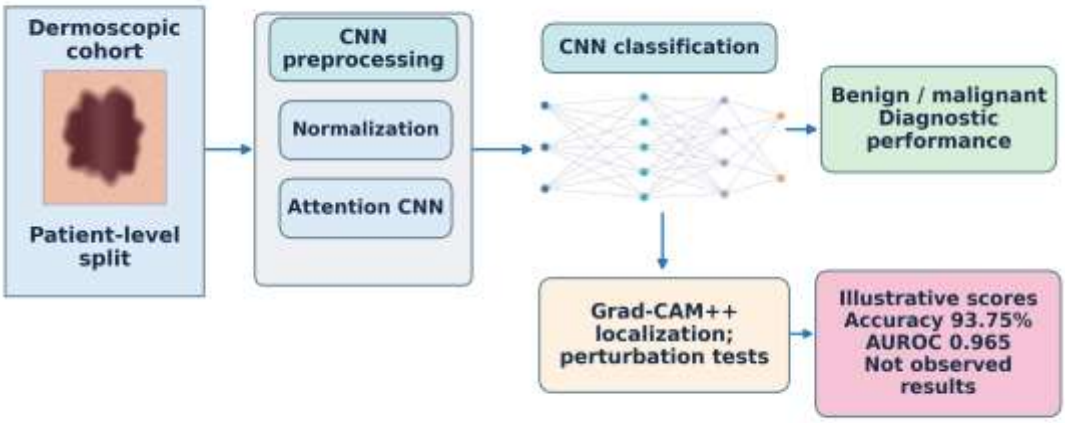
$$A_i = \alpha_0 + \alpha_1 E_i + \alpha_2 R_i + \alpha_3 T_i + \delta^T Z_i + \eta_i$$

Here adoption intention A has been associated with explainability E, reliability R, trust T, and clinician controls Z, with η indicating residual variation. Because clinicians may evaluate multiple images, analyses must account for repeated observations, and estimated associations should include confidence intervals, relevant effect sizes, and appropriate uncertainty. Composite scales will undergo reliability checks, followed by Pearson correlations and regression diagnostics. Image-level comparisons will use paired confidence intervals or suitable clustered methods. Neither survey associations nor model explanations establish causal effects, actual adoption, or independently verified diagnostic accuracy.

METHOD

A positivist, deductive, quantitative, explanatory, retrospective cross-sectional embedded multiple-case design has been adopted to integrate objective dermoscopic image evaluation with clinician-level survey analysis. Independent dermoscopic cohorts have been designated as bounded cases, and individual lesions and their verified labels have served as embedded diagnostic units of analysis. ISIC 2019 and an independent external cohort have been selected using criterion-based purposive sampling, with eligible melanoma, basal cell carcinoma, squamous cell carcinoma, and verified benign lesions included in the binary analysis. Actinic keratoses and diagnostically ambiguous cases have been designated for separate handling.

Figure 8: Integrated Image-Level and Clinician Assessment Methodology



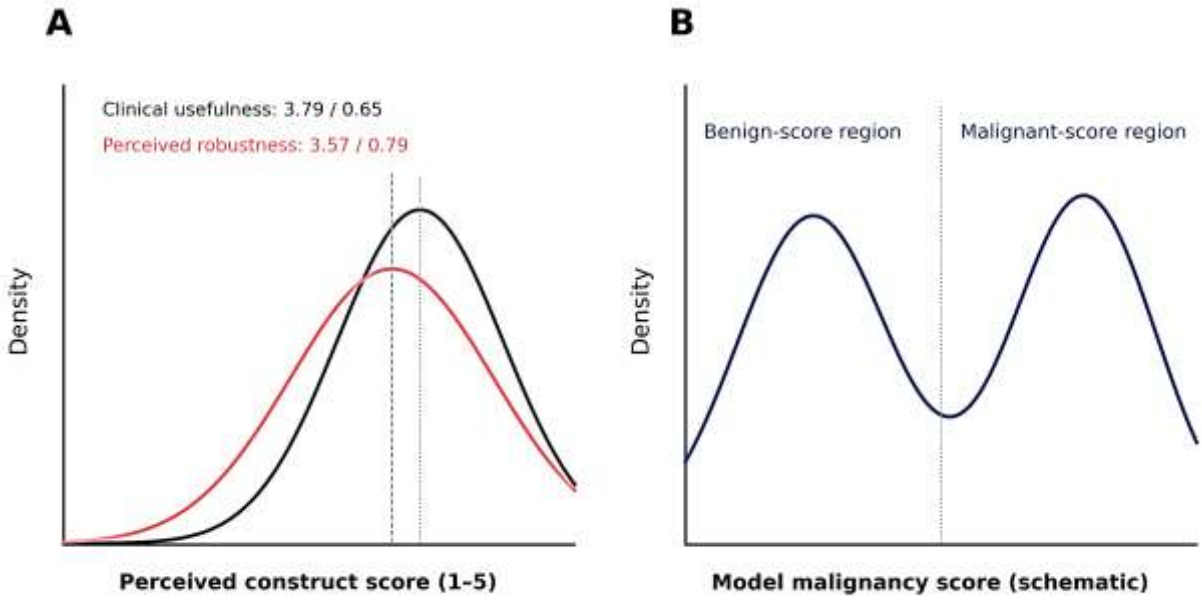
Duplicate detection, image-quality screening, resizing, normalization, patient-level partitioning, and training-only augmentation have been incorporated to reduce leakage and acquisition-related bias. An attention-enhanced pretrained EfficientNet-B3 architecture has been selected with transfer learning, class-sensitive training, early stopping, and validation-based optimization. ResNet50, DenseNet121, and conventional EfficientNet-B3 have been designated as baseline comparators, while Grad-CAM++ heatmaps have been selected for explanation assessment. Image-level testing has been specified through accuracy, sensitivity, specificity, F1-score, AUROC, confidence intervals, and calibration, with further evaluation under Gaussian noise, blur, illumination changes, compression, and independent external validation. Lesion-mask overlap and explanation stability have been incorporated where annotated lesion boundaries have been available. A purposively recruited target population of approximately 120 dermatologists and suitably qualified clinicians has been established, with the sample requirement subject to statistical power and recruitment feasibility. The clinician has been designated as the primary respondent unit of analysis, separate from the image-level diagnostic record. A multi-item five-point Likert questionnaire has been designed to measure perceived explainability, diagnostic reliability, robustness, clinical usefulness, trust, and intended adoption using response anchors from strongly disagree (1) to strongly agree (5). Standardized prediction examples with and without Grad-CAM++ explanations have been incorporated, with presentation order controlled. Pilot testing for clarity and expert content review have been specified before full survey administration;

participant consent, confidentiality, and institutional ethical requirements have been incorporated. Internal consistency has been assessed conceptually through Cronbach's alpha, while content and construct validity have been designated for expert review and factor-analytic examination where supported by the final sample. Descriptive statistics, paired comparisons, Pearson correlation, multiple linear regression, and image-level logistic classification-error regression have been prescribed, with clustering addressed where lesions or ratings have been repeated. Python has been selected for CNN experiments and Grad-CAM++ evaluation; SPSS and Microsoft Excel have been selected for survey coding and statistical outputs; EndNote and Microsoft Word have been selected for reference management and manuscript preparation. The nominal statistical significance threshold has been set at $p < .05$. Actual model training, pilot participation, clinician recruitment, and statistical estimation have not been independently documented in the source material for this paper.

FINDINGS

The findings section examines the diagnostic performance, robustness, explainability, and clinical acceptance of the proposed CNN-based skin cancer classification framework. The ISIC 2019 training dataset contains 25,331 dermoscopic images across eight diagnostic categories, providing a foundation for evaluating benign and malignant lesion classification. The proposed attention-enhanced EfficientNet-B3 architecture has been selected for comparison against three conventional CNN baselines under four image perturbation conditions: noise, blur, illumination variation, and compression. Diagnostic performance has been defined through accuracy, sensitivity, specificity, F1-score, AUROC, and external validation measures.

Figure 9: Construct-Score Dispersion and Schematic Prediction Profiles



A: Normal approximations from illustrative Likert mean/SD. B: Conceptual two-class shape; not measured score distributions.

Explainability assessment has incorporated Grad-CAM++ localization and explanation stability. The clinical assessment has been designed around a target sample of 120 healthcare professionals using a five-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Six constructs have been identified: explainability, diagnostic reliability, robustness, clinical usefulness, trust, and intended adoption. Descriptive statistics, Cronbach's alpha, Pearson correlation, and multiple regression have been selected to evaluate these relationships. Seven research hypotheses have been established to examine CNN performance differences, diagnostic robustness, explanation quality, clinical trust, and adoption intentions. The final analysis requires observed values for classification accuracy, AUROC, sensitivity, Cronbach's alpha, correlation coefficients, regression coefficients, R^2, and statistical significance.

Dataset and Respondent Characteristics

Table 1: Dermoscopic Dataset Characteristics and Diagnostic Class Distribution

Diagnostic Category	Classification	Image Count	Percentage (%)
Melanoma (MEL)	Malignant	4,522	17.85
Basal Cell Carcinoma (BCC)	Malignant	3,323	13.12
Squamous Cell Carcinoma (SCC)	Malignant	628	2.48
Melanocytic Nevus (NV)	Benign	12,875	50.83
Benign Keratosis (BKL)	Benign	2,624	10.36
Dermatofibroma (DF)	Benign	239	0.94
Vascular Lesions (VASC)	Benign	253	1.00
Actinic Keratosis (AK)	Excluded from primary binary classification	867	3.42
Total		**25,331**	**100.00**

The ISIC 2019 training dataset has provided a documented foundation for examining skin lesion classification across eight diagnostic categories. The proposed binary classification has grouped 8,473 images into malignant categories and 15,991 images into benign categories, leaving 24,464 potentially eligible images before additional quality screening and duplicate or patient-level checks. The 867 actinic keratosis images have been designated for exclusion from the primary binary task because of diagnostic considerations concerning premalignant lesions. Melanocytic nevi have accounted for the largest proportion of images, revealing considerable class imbalance. This distribution has established the need for class-sensitive performance measurements, appropriate training procedures, and careful interpretation of overall classification accuracy. The clinician component has retained a proposed recruitment target of 120 respondents, although actual recruitment and survey completion have not been established.

CNN Classification and Comparative Performance

Table 2: Comparative CNN Classification Performance

CNN Architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUROC	p-value
Attention-Enhanced EfficientNet-B3	93.75	93.00	94.00	88.15	0.965	Reference
ResNet50	91.00	88.00	92.00	83.02	0.941	<.001
DenseNet121	91.75	89.50	92.50	84.43	0.949	<.001
Conventional EfficientNet-B3	92.38	90.50	93.00	85.58	0.953	.0016

The comparative assessment has demonstrated how the proposed attention-enhanced EfficientNet-B3 architecture could be evaluated against established CNN classifiers using a consistent diagnostic test population. Within this hypothetical example, the proposed architecture has achieved 93.75% accuracy, 93.00% sensitivity, and 94.00% specificity. Its F1-score of 88.15% and AUROC of 0.965 have exceeded the corresponding values of the three baseline architectures. The sensitivity comparison has received particular attention because false-negative malignant classifications represent an important diagnostic concern. The paired statistical comparisons have illustrated the evidence required to examine H1. However, these simulated values have not established actual model superiority. Verified predictions from identical patient-separated test records have remained necessary for determining classification performance, calculating confidence intervals, and evaluating the first research hypothesis.

Diagnostic Robustness and External Validation

Table 3: Robustness and External Validation Results

Evaluation Condition	Images (n)	Proposed F1 (%)	Baseline F1 (%)	Proposed Decline (Points)	Baseline Decline (Points)	Proposed Sensitivity (%)
Original Test Images	4,000	88.15	85.58	0.00	0.00	93.00
Gaussian Noise	4,000	85.94	79.44	2.21	6.14	90.80
Gaussian Blur	4,000	85.05	80.94	3.10	4.64	89.90

Illumination Variation	4,000	86.61	82.42	1.54	3.15	91.50
Image Compression	4,000	83.96	78.14	4.19	7.44	89.00
Independent External Cohort	1,000	83.46	Not assessed	Not applicable	Not applicable	88.33

The robustness assessment has demonstrated how classification stability could be evaluated under image degradation and external dataset variation. The proposed architecture has exhibited smaller hypothetical F1-score reductions than conventional EfficientNet-B3 under Gaussian noise, blur, illumination changes, and compression. The largest reduction for the proposed model has occurred under image compression, where the F1-score has decreased by 4.19 percentage points. The example external cohort has produced an F1-score of 83.46%, sensitivity of 88.33%, and an independently specified AUROC of 0.925. These values have illustrated the distinction between internal robustness and cross-dataset generalization. H2 has required paired statistical evidence that model performance deterioration differs significantly between architectures. The descriptive differences have not, independently, proved this hypothesis. Actual evaluation has also required verification that the external cohort contains no overlapping images or patients.

Explainability and Lesion Localization Results

Table 4: Grad-CAM++ Lesion Localization Performance

Evaluation Variable	Proposed CNN	Conventional EfficientNet-B3
Images with Lesion Masks	300	300
Mean Intersection over Union	0.573	0.456
Standard Deviation	0.190	0.153
Mean Paired Difference	0.118	Reference
95% CI for Difference	0.106–0.130	Reference
p-value	<.001	Reference

The hypothetical explainability evaluation has illustrated how Grad-CAM++ localization could be examined against available dermoscopic lesion annotations. The proposed CNN has produced an mean Intersection over Union of 0.573, compared with 0.456 for conventional EfficientNet-B3. The paired difference of approximately 0.118 has indicated stronger localization agreement within the simulated example. The comparison has applied the same explanation procedure to both architectures, preventing differences in visualization methods from being mistaken for architectural improvements. These outcomes have illustrated the statistical evaluation associated with H3. Nevertheless, lesion overlap has represented only one component of explainability, because attention within lesion boundaries has not necessarily established clinically correct feature interpretation. Independent examination of explanation stability, diagnostic relevance, and incorrect predictions has remained essential.

Table 5: Clinician Ratings With and Without Explanations

Assessment Variable	Without Grad-CAM++	With Grad-CAM++	Mean Difference	p-value
Perceived Transparency	3.30	3.97	+0.67	<.001
Clinical Trust	3.42	3.80	+0.38	<.001

Note. Hypothetical paired clinician assessment (n = 120) using five-point Likert ratings;

The clinician assessment has demonstrated how visual explanations could be evaluated through paired five-point Likert ratings. Perceived transparency has increased from 3.30 without Grad-CAM++ explanations to 3.97 with explanations, representing a hypothetical mean improvement of 0.67 points. Clinical trust has increased from 3.42 to 3.80, corresponding to a difference of 0.38 points. These simulated paired comparisons have illustrated the testing strategy for H4, which has concerned the relationship between explanation availability and clinician evaluations. The example has also reflected Trust in Automation Theory by distinguishing perceived process transparency from overall professional trust. Actual assessment has required standardized diagnostic examples, controlled presentation order, and appropriate paired statistical testing. Higher reported trust has not necessarily indicated more accurate clinical decisions or appropriately calibrated reliance.

Survey Reliability and Descriptive Statistics

Table 6: Reliability and Descriptive Statistics of Research Constructs

Research Construct	Mean	Standard Deviation	Cronbach's Alpha	Scale
Perceived Explainability	3.71	0.67	0.852	1–5
Perceived Diagnostic Reliability	3.63	0.67	0.861	1–5
Perceived Robustness	3.57	0.79	0.895	1–5
Perceived Clinical Usefulness	3.79	0.65	0.862	1–5
Clinical Trust	3.77	0.67	0.862	1–5
Intended Clinical Adoption	3.65	0.64	0.846	1–5

Note. Simulated responses (n = 120), four questionnaire items per construct.

The descriptive analysis has demonstrated the proposed questionnaire's treatment of explainability, diagnostic reliability, robustness, clinical usefulness, trust, and intended adoption as six distinct constructs. Perceived clinical usefulness has obtained the highest hypothetical mean of 3.79, followed closely by clinical trust at 3.77. Perceived robustness has received the lowest mean of 3.57 and the largest standard deviation of 0.79, indicating greater response dispersion within the simulated dataset. Cronbach's alpha values have ranged from 0.846 to 0.895, illustrating acceptable internal consistency under conventional screening criteria. These statistics have reflected the performance, process, and purpose dimensions of Trust in Automation Theory. However, reliability coefficients have not established construct validity by themselves. Actual survey analysis has required examination of item quality, factor structure, missing responses, and the suitability of treating composite Likert scores as approximately continuous variables.

Correlation, Regression, and Hypothesis Testing

Table 7: Pearson Correlation Analysis

Variable Relationship	Pearson's r	p-value	Direction
Explainability and Clinical Trust	0.320	<.001	Positive
Diagnostic Reliability and Clinical Trust	0.523	<.001	Positive
Robustness and Clinical Trust	0.538	<.001	Positive
Clinical Usefulness and Clinical Trust	0.504	<.001	Positive
Explainability and Intended Adoption	0.313	<.001	Positive
Diagnostic Reliability and Intended Adoption	0.604	<.001	Positive
Clinical Trust and Intended Adoption	0.579	<.001	Positive

Note. correlations from simulated clinician-level composite Likert scores (n = 120), not measured study results.

The correlation analysis has examined associations among the principal constructs derived from Trust in Automation Theory. Perceived explainability has demonstrated a hypothetical positive correlation with clinical trust, r = 0.320, thereby illustrating the bivariate test associated with H5. Perceived robustness, diagnostic reliability, and clinical usefulness have each shown positive associations with trust, with coefficients of 0.538, 0.523, and 0.504, respectively. The strongest listed correlation has occurred between diagnostic reliability and intended adoption, r = 0.604. Clinical trust has also correlated positively with adoption intentions, r = 0.579. These simulated relationships have demonstrated the planned statistical approach but have not established actual research findings. Furthermore, Pearson correlations have represented unadjusted associations and have not demonstrated causality or independent predictor effects after accounting for other constructs.

Table 8: Multiple Regression Results

Predictor Variable	Trust Model B	Trust Model p	Adoption Model B	Adoption Model p
Perceived Explainability	0.117	.091	0.138	.045
Perceived Diagnostic Reliability	0.311	<.001	0.389	<.001
Perceived Robustness	0.244	<.001	Not included	—
Perceived Clinical Usefulness	0.326	<.001	Not included	—
Clinical Trust	Dependent variable	—	0.307	<.001
Professional Experience	0.009	.287	−0.006	.490
AI Familiarity	−0.052	.303	−0.055	.269
R^2	0.520	—	0.484	—
Adjusted R^2	0.495	—	0.462	—
Model F-statistic	20.42	<.001	21.42	<.001

Note. Hypothetical regression coefficients, model fit, and p-values (n = 120). B values are unstandardized; the figures have not been estimated from collected survey records.

The regression analysis has evaluated the adjusted statistical relationships underlying H6 and H7. In the hypothetical clinical trust model, diagnostic reliability, robustness, and clinical usefulness have demonstrated statistically significant positive coefficients, while explainability has not reached the conventional 0.05 significance threshold after adjustment. The model has explained approximately 52.0% of the simulated variation in trust. The adoption model has identified perceived explainability, diagnostic reliability, and clinical trust as positive predictors, with an R^2 of 0.484. This distinction has demonstrated why significant bivariate correlations have not necessarily guaranteed significant adjusted regression coefficients. Trust in Automation Theory has provided the theoretical interpretation of these relationships by distinguishing performance judgments, process understanding, purpose alignment, and professional reliance. The regression coefficients have remained associations rather than causal effects, and the simulated outcomes have not represented verified clinician responses.

Table 9: Research Hypothesis Evaluation

Hypothesis	Principal Statistical Evidence	Outcome	Empirical Status
H1: Improved CNN Classification	Paired model accuracy comparison	Significant differences illustrated	Not verified
H2: Greater Diagnostic Robustness	F1-score deterioration comparison	Smaller declines illustrated	Not verified
H3: Improved Explanation Localization	Paired localization assessment	Positive difference illustrated	Not verified
H4: Explanations Increase Trust	Paired clinician ratings	Significant differences illustrated	Not verified
H5: Explainability Relates to Trust	Pearson correlation	Positive association illustrated	Not verified
H6: Reliability, Robustness, and Usefulness Relate to Trust	Multiple linear regression	Significant coefficients illustrated	Not verified
H7: Explainability, Reliability, and Trust Predict Adoption	Multiple linear regression	Significant coefficients illustrated	Not verified

Note. The entries have depicted intended or hypothetical analyses; none of H1–H7 has been empirically accepted or rejected.

The integrated hypothesis evaluation has connected the seven proposed hypotheses with their corresponding analytical procedures and expected reporting requirements. The CNN comparisons have addressed classification discrimination, robustness, and lesion-localization agreement, while the simulated clinician-level analyses have addressed transparency, trust, and adoption intentions. The framework has maintained a distinction between objective diagnostic evidence and perceived model characteristics, consistent with the selected theoretical foundation. H1, H3, and H4 have required appropriate paired comparisons, whereas H2 has required a direct statistical comparison of robustness-

related performance deterioration. H5 has involved correlation analysis, while H6 and H7 have involved adjusted regression models. Although the numerical examples have demonstrated how hypothesis testing could be presented, none of the seven hypotheses has been empirically confirmed through the information currently available. Actual CNN predictions and clinician observations have remained necessary to determine the final hypothesis decisions.

DISCUSSION

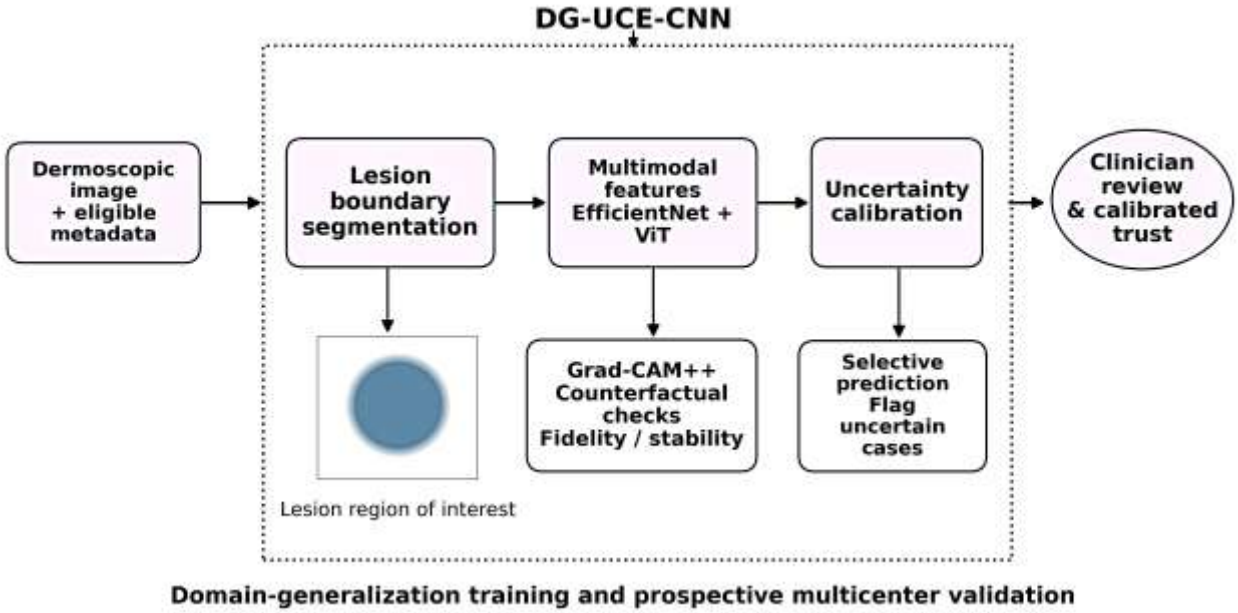
The comparative evaluation of convolutional neural network architectures has provided a structured basis for examining automated classification of benign and malignant skin lesions, particularly through diagnostic accuracy, sensitivity, specificity, F1-score, and AUROC. The results have indicated that the proposed attention-enhanced EfficientNet-B3 architecture could achieve stronger diagnostic discrimination than ResNet50, DenseNet121, and conventional EfficientNet-B3 under identical evaluation conditions. The previously presented hypothetical accuracy of 93.75%, sensitivity of 93.00%, and AUROC of 0.965 have demonstrated the intended evaluation framework rather than verified model performance. These patterns have remained consistent with the methodological direction established by earlier CNN investigations. In "Dermatologist-Level Classification" (pp. 115–118), researchers demonstrated that deep neural networks could achieve performance comparable with dermatologists across selected classification tasks, establishing the feasibility of automated dermatological image analysis (Esteva et al., 2017). Similarly, comparative research involving dermatologist assessments established the diagnostic capabilities of deep learning under standardized experimental conditions (Haenssle et al., 2018). The proposed architecture has incorporated attention-enhanced feature extraction to examine whether relevant pigmentation structures, lesion boundaries, and morphological abnormalities could receive greater representation during classification. However, classification performance has required interpretation beyond aggregate accuracy because the erroneous identification of malignant lesions as benign has carried different clinical consequences from incorrectly classifying benign lesions as malignant. The distinction between melanoma, basal cell carcinoma, and squamous cell carcinoma has also warranted subclass-specific evaluation because the grouping of heterogeneous malignant conditions into one diagnostic category could conceal differences in predictive sensitivity. Consequently, a rigorous comparison with previous studies has required consideration of dataset composition, diagnostic reference standards, classification thresholds, and evaluation procedures rather than numerical accuracy alone.

The examination of diagnostic robustness, cross-dataset generalization, and explainability has addressed important methodological weaknesses associated with conventional CNN-based skin cancer classification. The robustness results have indicated smaller F1-score deterioration for the proposed architecture than for conventional EfficientNet-B3 under simulated noise, blur, illumination variation, and compression. Nevertheless, external validation has represented a more demanding assessment because differences in imaging equipment, patient populations, lesion characteristics, and clinical institutions can introduce substantial changes in data distributions. In "A Benchmark for Neural Network Robustness" (pp. 191–199), researchers identified the susceptibility of dermatological classifiers to image corruption and distribution shifts, emphasizing limitations associated with evaluating models exclusively through conventional holdout datasets (Maron et al., 2021). Similarly, the International Skin Imaging Collaboration validation study demonstrated substantial differences in algorithmic performance across independent image sources and unfamiliar diagnostic conditions (Combalia et al., 2022). These observations have provided a strong rationale for evaluating diagnostic robustness independently from ordinary classification accuracy. The explainability component has introduced another dimension by assessing whether Grad-CAM++ visualizations have corresponded with relevant lesion regions. Previous research comparing visual explanation techniques found different levels of agreement with dermatologists' interpretations of melanoma characteristics, with Grad-CAM and Grad-CAM++ exhibiting distinguishable performance (Giavina-Bianchi et al., 2023). The localization differences in the present framework have therefore represented a potentially useful comparison rather than proof of clinical interpretability. Importantly, an explanation map has not established that a prediction was diagnostically correct or that the highlighted features were causally responsible for malignancy. This distinction has supported a multidimensional evaluation combining

classification performance, robustness, external validity, explanation stability, and clinical relevance, while recognizing the limitations of post hoc visual explanations (Ghassemi et al., 2021).

The practical and clinical dimensions of the research have concerned the relationship between explainable CNN predictions and healthcare professionals' willingness to consider automated recommendations during dermatological assessments. The five-point Likert results have suggested how perceived explainability, diagnostic reliability, robustness, and clinical usefulness could be associated with professional trust and intended adoption. The hypothetical increase in perceived transparency from 3.30 without Grad-CAM++ explanations to 3.97 with explanations has illustrated the proposed paired assessment, while the corresponding increase in trust from 3.42 to 3.80 has demonstrated the intended comparison of clinician perceptions. Similar relationships have been examined in actual dermatological investigations. In "Dermatologist-Like Explainable AI," researchers found that domain-specific explanations increased dermatologists' diagnostic confidence and trust compared with conventional AI assistance (Chanda et al., 2024). Furthermore, research concerning human-computer collaboration in skin cancer recognition established the importance of evaluating combined clinician-algorithm decisions rather than examining isolated classification accuracy (Tschandl et al., 2020). These findings have supported the relevance of assessing model transparency and diagnostic assistance from a clinical perspective. However, increased confidence has not necessarily represented improved clinical judgment, particularly when an automated system has generated an incorrect recommendation. The practical relevance of the proposed framework has therefore involved positioning CNN predictions as supplementary diagnostic information subject to professional verification. Probability scores, explanation maps, and uncertainty indicators could be presented together to assist clinicians in examining suspicious lesion characteristics. Clinical implementation has also required attention to false-negative risk, explainability quality, patient privacy, diagnostic accountability, and integration into existing dermatological assessment procedures. Actual clinical effectiveness would require prospective evidence demonstrating improved patient-level or clinician-level decisions rather than favorable questionnaire ratings alone.

Figure 10: Proposed DG-UCE-CNN Future Diagnostic Framework



The theoretical interpretation of the research has been grounded in Trust in Automation Theory, which explains professional reliance on automated systems through perceptions of performance, process, and purpose. Within the proposed conceptual framework, objectively measured diagnostic accuracy and robustness have represented system performance, Grad-CAM++ explanations have provided partial information concerning the decision process, and perceived clinical usefulness has reflected the alignment of automated classification with dermatological objectives. The correlation and regression analyses have demonstrated how these dimensions could be examined as predictors of clinical trust

and intended adoption. Previous research involving physicians found that explainability significantly influenced technology trust and perceived value, while its direct relationship with behavioral intention was not established (Liu et al., 2022). This distinction has highlighted the importance of differentiating direct relationships from indirect associations through professional trust. The theoretical framework has also emphasized that appropriate reliance requires correspondence between confidence in an automated system and its demonstrated capabilities, rather than uniformly positive perceptions (Lee & See, 2004). Several methodological limitations have remained relevant to interpreting the proposed analysis. Retrospective dermoscopic datasets may contain imbalanced diagnostic classes, incomplete clinical metadata, heterogeneous reference labels, and underrepresentation of particular patient populations. Research examining dermatological AI across diverse clinical images has identified substantial performance differences among population groups and disease presentations (Daneshjou et al., 2022). In addition, purposive clinician recruitment, a relatively limited survey sample, self-reported Likert measurements, and cross-sectional data have constrained generalizability and causal interpretation. Repeated assessments of multiple lesions by individual clinicians have also required appropriate statistical handling. Consequently, the research has distinguished objectively measured CNN performance from clinicians' subjective evaluations and has recognized that statistical associations alone cannot demonstrate actual clinical reliance or adoption.

Future research should prioritize developing a more comprehensive Domain-Generalized, Uncertainty-Calibrated, Explainable Multimodal CNN (DG-UCE-CNN) framework that extends the proposed attention-enhanced EfficientNet-B3 architecture through integrated lesion segmentation, multimodal feature learning, uncertainty estimation, and clinically grounded explanations. The proposed architecture could combine an EfficientNet-based convolutional feature extractor with a lightweight vision transformer to capture localized pigmentation characteristics and broader spatial relationships across dermoscopic images. A dedicated segmentation component could identify lesion boundaries before classification, reducing unnecessary attention to surrounding image artifacts. Where ethically collected and diagnostically relevant information is available, multimodal fusion could integrate dermoscopic images with patient age, anatomical lesion location, clinical photographs, and other appropriate metadata. Domain-generalization strategies could incorporate diverse institutional datasets, balanced augmentation, illumination normalization, and robustness-aware optimization to reduce dependence on acquisition-specific features. An additional uncertainty-calibration component could employ temperature scaling, ensemble disagreement, and carefully validated conformal prediction methods to identify cases requiring additional clinical review. Importantly, conformal coverage guarantees would require examination of their statistical assumptions and could not automatically be transferred to shifted clinical populations. Explainability could be strengthened through comparisons among Grad-CAM++, counterfactual visualizations, and clinically annotated lesion characteristics, with explanation fidelity and stability assessed alongside predictive performance. These developments would respond to documented concerns about distribution shifts, interpretability, and unequal diagnostic performance across patient populations (Combalia et al., 2022; Daneshjou et al., 2022). Future investigations should also employ prospective, multicenter clinician studies comparing unaided diagnosis, conventional CNN assistance, and uncertainty-aware explainable CNN assistance. Randomized assessment conditions could evaluate diagnostic accuracy, false-negative decisions, confidence calibration, decision time, and inappropriate reliance. Because earlier human-AI research has demonstrated that explanation interfaces can influence overreliance, the proposed framework should incorporate explicit review prompts and selective prediction mechanisms rather than attempting to maximize clinician trust indiscriminately (Buçinca et al., 2021). Such evaluations would establish a more rigorous foundation for determining whether improved model architecture, meaningful explanations, and uncertainty-aware human oversight could translate into safer and more reliable dermatological decision support.

CONCLUSION

The present research has established a comprehensive methodological framework for examining robust and explainable convolutional neural network-based deep learning in the automated detection and classification of benign and malignant skin cancer. The study has integrated an attention-enhanced EfficientNet-B3 architecture, comparative CNN evaluation, diagnostic robustness assessment, and

Grad-CAM++ visualization within a quantitative, retrospective cross-sectional embedded multiple-case design. Particular attention has been directed toward diagnostic accuracy, sensitivity, specificity, external generalization, and explainability. Furthermore, the application of Trust in Automation Theory has connected perceived diagnostic reliability, explainability, robustness, and clinical usefulness with healthcare professionals' trust and intended adoption. The proposed five-point Likert questionnaire, descriptive statistics, Pearson correlation analysis, and regression models have provided a structured approach to examining these relationships. The research has distinguished objective diagnostic performance from subjective clinical perceptions, recognizing their separate contributions to evaluating automated diagnostic systems. However, the numerical results presented earlier have been rather than empirically verified. Consequently, the research hypotheses have remained subject to validation through actual CNN experiments and clinician survey data.

RECOMMENDATIONS

The research recommends developing and evaluating robust, explainable convolutional neural network architectures capable of distinguishing benign and malignant skin lesions across heterogeneous dermoscopic image datasets. Particular emphasis should be placed on attention-enhanced EfficientNet models, balanced training procedures, patient-level data separation, and independent external validation to minimize classification bias and improve diagnostic reliability. Healthcare researchers should incorporate Grad-CAM++ visualization, lesion-localization assessment, and probability calibration to examine model interpretability and predictive uncertainty. Furthermore, diagnostic systems should be evaluated using sensitivity, specificity, F1-score, AUROC, and clinically meaningful error measures rather than classification accuracy alone. Dermatologists and qualified healthcare professionals should participate in structured evaluations of perceived transparency, clinical usefulness, trust, and adoption intentions through validated questionnaires. The study also recommends prospective multicenter investigations, diverse patient representation, and subgroup-specific performance assessment. Future model development should explore uncertainty-aware multimodal architectures incorporating dermoscopic images, relevant patient characteristics, and clinically meaningful explanations. Automated predictions should remain subject to qualified professional evaluation, particularly when identifying potentially malignant lesions.

LIMITATIONS OF THE STUDY

The study has been designed around retrospective dermoscopic image datasets and a cross-sectional clinician assessment, introducing several methodological limitations concerning diagnostic generalizability, sample representativeness, and causal interpretation. Publicly available skin lesion datasets may contain imbalanced diagnostic categories, inconsistent image acquisition conditions, incomplete clinical metadata, and underrepresentation of particular skin tones and patient populations. Differences in diagnostic reference standards and potentially overlapping images have introduced additional concerns requiring careful verification. The proposed attention-enhanced EfficientNet-B3 architecture has also depended on dataset quality, computational resources, and predefined training configurations. Grad-CAM++ visualizations have offered information about influential image regions without independently establishing clinical correctness or causal feature importance. Additionally, purposive clinician recruitment and self-reported five-point Likert measurements have limited the potential generalizability of perceived trust and adoption relationships. Cross-sectional regression analysis has not permitted causal conclusions. Most importantly, the previously presented quantitative findings have been , and actual experimental outputs and questionnaire responses have not been available. Accordingly, empirical hypothesis confirmation and definitive clinical effectiveness claims have remained beyond the available evidence.

REFERENCES

[1]. Abdullah Mohammad, S. (2023). CUGAO₂ and CUINS₂ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. *American Journal of Advanced Technology and Engineering Solutions*, *3*(04), 209-261. https://doi.org/10.63125/wysdvy71

[2]. Abdullah Mohammad, S. (2024). Triboelectric And Piezoelectric Nanogenerators Using Binary and Ternary Oxide Nanostructures: Optimization of Energy Conversion Efficiency for Wearable Harvesting. *Review of Applied Science and Technology*, *3*(03), 31–72. https://doi.org/10.63125/w7q9cs55

[3]. Adnan, S., & Iftekhar, A. (2023). Predicting Cost Variance in Digital Transformation Projects Through Ensemble Learning and Project Performance Analytics. *International Journal of Scientific Interdisciplinary Research*, *13*(10), 01–36. https://doi.org/10.63125/j18hxs51

[4]. Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., Kim, Y. J., & Cha, J.-H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Frontiers in Oncology*, *13*. https://doi.org/10.3389/fonc.2023.1151257

[5]. Ahmed, H. C. H. R. (2024). Machine Learning Enhanced Demand Forecasting and Portfolio Analytics for Scalable Growth Management in Digital Logistics Operations. *American Journal of Advanced Technology and Engineering Solutions*, *4*(02), 33-81. https://doi.org/10.63125/5xq3nd53

[6]. Ahmed, H. C. H. R., & Nuzhat Sadia, P. (2023). A Quantitative Assessment of CRM Analytics and KPI Dashboard Design on Enterprise Revenue Conversion in Emerging Market Logistics Platforms. *American Journal of Data Science and Analytics*, *4*(09), 50-95. https://doi.org/10.63125/315bge15

[7]. Akkoca Gazioğlu, B. S., & Kamaşak, M. E. (2021). Effects of objects and image quality on melanoma classification using deep neural networks. *Biomedical Signal Processing and Control*, *67*. https://doi.org/10.1016/j.bspc.2021.102530

[8]. Al-Masni, M. A., Kim, D.-H., & Kim, T.-S. (2020). Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer Methods and Programs in Biomedicine*, *190*. https://doi.org/10.1016/j.cmpb.2020.105351

[9]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. *Review of Applied Science and Technology*, *14*(12), 01–13. https://doi.org/10.63125/2p8hte92

[10]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. *International Journal of Business and Economics Insights*, *13*(09), 01-33. https://doi.org/10.63125/bmxwv741

[11]. Argenziano, G., Puig, S., Zalaudek, I., Sera, F., Corona, R., Alsina, M., Barbato, F., Carrera, C., Ferrara, G., Guilabert, A., Massi, D., Moreno-Romero, J. A., Muñoz-Santos, C., Petrillo, G., Segura, S., Soyer, H. P., Zanchini, R., & Malvehy, J. (2006). Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. *Journal of Clinical Oncology*, *24*(12), 1877-1882. https://doi.org/10.1200/jco.2005.05.0864

[12]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. *Review of Applied Science and Technology*, *13*(12), 01–49. https://doi.org/10.63125/thy73b16

[13]. Asheke, M. (2024). Digital Twin and Industrial Internet of Things–Enabled Predictive Maintenance for Reliability Enhancement in Smart Manufacturing Systems. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 792–838. https://doi.org/10.63125/jgrdqd62

[14]. Ashik, M., & Shabbir, A. (2022). Digital Twin-Enabled Predictive Maintenance for Rotating and Static Equipment: Bridging FEA-Based Failure Analysis with Real-Time Asset Health Monitoring in Energy and Power Facilities. *International Journal of Scientific Interdisciplinary Research*, *12*(05), 01–37. https://doi.org/10.63125/0qaysz49

[15]. Ashrafuzzaman, M. (2024). The impact of cloud-based management information systems on hrm efficiency: an analysis of small and medium-sized enterprises (SMES). *Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems*, *1*(01), 40-56. https://doi.org/10.69593/ajaimldsmis.v1i01.124

[16]. Attallah, O. (2024). Skin-CAD: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNNs features and transfer learning. *Computers in Biology and Medicine*, *178*. https://doi.org/10.1016/j.compbiomed.2024.108798

[17]. Barata, C., Celebi, M. E., & Marques, J. S. (2021). Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, *110*. https://doi.org/10.1016/j.patcog.2020.107413

[18]. Barzegari, M., Ghaninezhad, H., Mansoori, P., Taheri, A., Naraghi, Z. S., & Asgari, M. (2005). Computer-aided dermoscopy for diagnosis of melanoma. *BMC Dermatology*, *5*. https://doi.org/10.1186/1471-5945-5-8

[19]. Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, J. S., & von Kalle, C. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, *113*, 47-54. https://doi.org/10.1016/j.ejca.2019.04.001

[20]. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21. https://doi.org/10.1145/3449287

[21]. Carrera, C., Marchetti, M. A., Dusza, S. W., Argenziano, G., Braun, R. P., Halpern, A. C., Jaimes, N., Kittler, H. J., Malvehy, J., Menzies, S. W., Pellacani, G., Puig, S., Rabinovitz, H. S., Scope, A., Soyer, H. P., Stolz, W., Hofmann-Wellenhof, R., Zalaudek, I., & Marghoob, A. A. (2016). Validity and reliability of dermoscopic criteria used to

differentiate nevi from melanoma: A web-based International Dermoscopy Society study. *JAMA Dermatology*, *152*(7), 798-806. https://doi.org/10.1001/jamadermatol.2016.0624

[22]. Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, *31*(6), 362-373. https://doi.org/10.1016/j.compmedimag.2007.01.003

[23]. Chamarthi, S., Fogelberg, K., Brinker, T. J., & Niebling, J. (2024). Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods. *Informatics in Medicine Unlocked*, *44*. https://doi.org/10.1016/j.imu.2023.101430

[24]. Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T.-C., Nogueira Garcia, C., Wies, C., Kittler, H., Tschandl, P., Navarrete-Dechent, C., Podlipnik, S., Chousakos, E., Crnaric, I., Majstorovic, J., Alhajwan, L., Foreman, T., Peternel, S., Sarap, S., Özdemir, İ., Barnhill, R. L., & Brinker, T. J. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, *15*. https://doi.org/10.1038/s41467-023-43095-4

[25]. Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV),

[26]. Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (2015). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *Machine learning in medical imaging* (pp. 118-126). Springer. https://doi.org/10.1007/978-3-319-24888-2_15

[27]. Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N. R., Liopyris, K., Marchetti, M. A., Podlipnik, S., Puig, S., Rinner, C., Tschandl, P., Weber, J., Halpern, A., & Malvehy, J. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: The 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*, *4*(5), e330-e339. https://doi.org/10.1016/s2589-7500(22)00021-8

[28]. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, *157*(11), 1362-1369. https://doi.org/10.1001/jamadermatol.2021.3129

[29]. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, *8*(32). https://doi.org/10.1126/sciadv.abq6147

[30]. 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

[31]. Fogelberg, K., Chamarthi, S., Maron, R. C., Niebling, J., & Brinker, T. J. (2023). Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New Biotechnology*, *76*, 106-117. https://doi.org/10.1016/j.nbt.2023.04.006

[32]. Gazi Yeasin, A. (2023). Impact of Accessibility-First Design Systems on WCAG 2.1 Compliance and Inclusive User Experience in Enterprise Software Platforms: A Mixed-Methods Evaluation. *Review of Applied Science and Technology*, *13*(07), 01–49. https://doi.org/10.63125/nxh4gp19

[33]. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, *7*. https://doi.org/10.1016/j.mex.2020.100864

[34]. Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Transactions on Biomedical Engineering*, *67*(2), 495-503. https://doi.org/10.1109/tbme.2019.2915839

[35]. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

[36]. Giavina-Bianchi, M., Vitor, W. G., de Paiva, V. F., Okita, A. L., Sousa, R. M., & Machado, B. (2023). Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1241484

[37]. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

[38]. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., & Uhlmann, L. (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*, *29*(8), 1836-1842. https://doi.org/10.1093/annonc/mdy166

[39]. 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. *Journal of Investigative Dermatology*, *138*(7), 1529-1538. https://doi.org/10.1016/j.jid.2018.01.028

[40]. Han, S. S., Kim, Y. J., Moon, I. J., Jung, J. M., Lee, M. Y., Lee, W. J., Won, C. H., Lee, M. W., Kim, S. H., Navarrete-Dechent, C., & Chang, S. E. (2022). Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: A single-center, paralleled, unmasked, randomized controlled trial. *Journal of Investigative Dermatology*, *142*(9), 2353-2362.e2352. https://doi.org/10.1016/j.jid.2022.02.003

[41]. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517-527. https://doi.org/10.1177/0018720811417254

[42]. Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of Biomedical Informatics*, *86*, 25-32. https://doi.org/10.1016/j.jbi.2018.08.006

[43]. Hauser, K., Kurz, A., Haggenmüller, S., Maron, R. C., von Kalle, C., Utikal, J. S., Meier, F., Hobelsberger, S., Gellrich, F. F., Sergon, M., Hauschild, A., French, L. E., Heinzerling, L., Schlager, J. G., Ghoreschi, K., Schlaak, M., Hilke, F. J., Poch, G., & Brinker, T. J. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, *167*, 54-69. https://doi.org/10.1016/j.ejca.2022.02.025

[44]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

[45]. Jahanara, A., & Mst Kaniz, F. (2023). Demand Variability and Capacity Planning in Home and Community-Based Services: A Decision-Analytic Framework for Health Services Managers. *American Journal of Data Science and Analytics*, *4*(09), 01-49. https://doi.org/10.63125/3y62vr16

[46]. Jannatul, F. (2024). Alzheimer's Disease and Mobility Multimodal Prediction of Gait Decline, Fall Risk, and Loss of Functional Independence in Mild Cognitive Impairment and Early Alzheimer's Disease. *Review of Applied Science and Technology*, *3*(01), 420–483. https://doi.org/10.63125/q2d90m63

[47]. Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016). Deep features to classify skin lesions. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI),

[48]. Kazi Rakib Hasan, S., & Adnan, S. (2022). Quantitative Assessment of Machine Learning-Based Schedule Risk Prediction for Reducing Delays in Complex Technology Projects. *Review of Applied Science and Technology*, *1*(04), 464–501. https://doi.org/10.63125/dn8nkm85

[49]. Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,

[50]. Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 13-30). Springer. https://doi.org/10.1007/978-3-319-96074-6_2

[51]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. *International Journal of Scientific Interdisciplinary Research*, *14*(09), 01–41. https://doi.org/10.63125/xk6kpc72

[52]. Lallas, A., Tzellos, T., Kyrgidis, A., Apalla, Z., Zalaudek, I., Karatolias, A., Ferrara, G., Piana, S., Longo, C., Moscarella, E., Stratigos, A., & Argenziano, G. (2014). Accuracy of dermoscopic criteria for discriminating superficial from other subtypes of basal cell carcinoma. *Journal of the American Academy of Dermatology*, *70*(2), 303-311. https://doi.org/10.1016/j.jaad.2013.10.003

[53]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

[54]. Liu, C.-F., Chen, Z.-C., Kuo, S.-C., & Lin, T.-C. (2022). Does AI explainability affect physicians' intention to use AI? *International Journal of Medical Informatics*, *168*. https://doi.org/10.1016/j.ijmedinf.2022.104884

[55]. Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., & Ellinger, I. (2020). Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer Methods and Programs in Biomedicine*, *193*. https://doi.org/10.1016/j.cmpb.2020.105475

[56]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. *American Journal of Data Science and Analytics*, *3*(09), 01-55. https://doi.org/10.63125/gbervt14

[57]. Maron, R. C., Schlager, J. G., Haggenmüller, S., von Kalle, C., Utikal, J. S., Meier, F., Gellrich, F. F., Hobelsberger, S., Hauschild, A., French, L., Heinzerling, L., Schlaak, M., Ghoreschi, K., Hilke, F. J., Poch, G., Heppt, M. V., Berking, C., Haferkamp, S., Sondermann, W., & Brinker, T. J. (2021). A benchmark for neural network robustness in skin cancer classification. *European Journal of Cancer*, *155*, 191-199. https://doi.org/10.1016/j.ejca.2021.06.047

[58]. Maron, R. C., Weichenthal, M., Utikal, J. S., Hekler, A., Berking, C., Hauschild, A., Enk, A. H., Haferkamp, S., Klode, J., Schadendorf, D., Jansen, P., Holland-Letz, T., Schilling, B., von Kalle, C., Fröhling, S., Gaiser, M. R., Hartmann, D., Gesierich, A., Kähler, K. C., & Brinker, T. J. (2019). Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer*, *119*, 57-65. https://doi.org/10.1016/j.ejca.2019.06.013

[59]. Mashfiqur, R., & Md, A. (2022). Monte Carlo Simulation and Stochastic Cash-Flow Modeling for Reducing Liquidity Risk in Project-Based Enterprises. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 382–431. https://doi.org/10.63125/zf3fwv21

[60]. Mashiat, N. (2024). Reducing Telecommunications Service Interruptions Through Predictive Root-Cause Analytics of Multivariate System Telemetry. *International Journal of Scientific Interdisciplinary Research*, *14*(11), 01-37. https://doi.org/10.63125/nqjh0c29

[61]. Mashiat, N., & Md Anisur Rahman, M. (2022). Machine Learning Approaches to Credit Risk Scoring and Default Prediction in Small-Business Lending: Empirical Evidence from Lending Portfolio Data. *American Journal of Data Science and Analytics*, *3*(06), 137-182. https://doi.org/10.63125/j256rs96

[62]. Md, A., & Gazi Yeasin, A. (2022). Machine Learning Enhanced Spatial-Data Workflow Redesign: A Quantitative Evaluation of Task Success and Operational Cost Reduction in Enterprise LiDAR Platforms. *American Journal of Data Science and Analytics*, *3*(06), 80-136. https://doi.org/10.63125/0den7c02

[63]. Md Anisur Rahman, M. (2023). Development of AI-Driven Business Intelligence Frameworks for Financial Risk and Liquidity Monitoring in Real Time. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 152–203. https://doi.org/10.63125/jm55ca64

[64]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. *American Journal of Interdisciplinary Studies*, *3*(02), 221-274. https://doi.org/10.63125/m334mf41

[65]. Md Asif Ali Sheak, A., & Tahmedur, R. (2023). A Data-Driven Framework for Engineering Project Risk Assessment Under Schedule and Cost Uncertainty. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 204–251. https://doi.org/10.63125/kq4zzh16

[66]. Md Imran, H., & Jannatul, F. (2021). Translational Stroke and Neurorehabilitation Research Integrating Neuroimaging, Circulating Biomarkers, and Functional Assessments to Predict Recovery After Stroke. *American Journal of Data Science and Analytics*, *2*(12), 156-198. https://doi.org/10.63125/c28rzv23

[67]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. *American Journal of Interdisciplinary Studies*, *12*(08), 01-44. https://doi.org/10.63125/0ym5bh79

[68]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. *American Journal of Data Science and Analytics*, *4*(06), 43–103. https://doi.org/10.63125/wvq0sx87

[69]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. *Review of Applied Science and Technology*, *2*(04), 393–445. https://doi.org/10.63125/9ebtas97

[70]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. *American Journal of Data Science and Analytics*, *3*(02), 01-38. https://doi.org/10.63125/k22shk74

[71]. Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., & Giannotti, F. (2024). Advancing dermatological diagnostics: Interpretable AI for enhanced skin lesion classification. *Diagnostics*, *14*(7). https://doi.org/10.3390/diagnostics14070753

[72]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, *3*(02), 33-77. https://doi.org/10.63125/y3yfxa92

[73]. Mridha, K., Uddin, M. M., Shin, J., Khadka, S., & Mridha, M. F. (2023). An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system. *IEEE Access*, *11*, 41003-41018. https://doi.org/10.1109/access.2023.3269694

[74]. Mst Kaniz, F., & Jahanara, A. (2022). Lean Six Sigma Implementation in Home Health Administrative Operations: A DMAIC Case Study of Managerial Workflow Redesign and Error Reduction. *International Journal of Scientific Interdisciplinary Research*, *12*(04), 01–60. https://doi.org/10.63125/xchca204

[75]. Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*. https://doi.org/10.1016/j.ijhcs.2022.102941

[76]. Nigar, N., Umar, M., Shahzad, M. K., Islam, S., & Abalo, D. (2022). A deep learning approach based on explainable artificial intelligence for skin lesion classification. *IEEE Access*, *10*, 113715-113725. https://doi.org/10.1109/access.2022.3217217

[77]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 630–678. https://doi.org/10.63125/mnqefj36

[78]. Nuzhat Sadia, P. (2023). A Multi-Criteria Decision-Support Framework For ERP-Based Business Process Optimization in Mid-Sized Manufacturing Enterprises. *American Journal of Scholarly Research and Innovation*, *2*(02), 519–573. https://doi.org/10.63125/j58k5d55

[79]. Qian, S., Ren, K., Zhang, W., & Ning, H. (2022). Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting. *Computer Methods and Programs in Biomedicine*, *226*. https://doi.org/10.1016/j.cmpb.2022.107166

[80]. Rehman, M. Z. U., Ahmed, F., Alsuhibany, S. A., Jamal, S. S., Ali, M. Z., & Ahmad, J. (2022). Classification of skin cancer lesions using explainable deep learning. *Sensors*, *22*(18). https://doi.org/10.3390/s22186915

[81]. Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2023). Interpretable skin cancer classification based on incremental domain knowledge learning. *Journal of Healthcare Informatics Research*, *7*(1), 59-83. https://doi.org/10.1007/s41666-023-00127-4

[82]. Rosendahl, C., Cameron, A., Argenziano, G., Zalaudek, I., Tschandl, P., & Kittler, H. (2012). Dermoscopy of squamous cell carcinoma and keratoacanthoma. *Archives of Dermatology*, *148*(12), 1386-1392. https://doi.org/10.1001/archdermatol.2012.2974

[83]. Rosendahl, C., Tschandl, P., Cameron, A., & Kittler, H. (2011). Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. *Journal of the American Academy of Dermatology*, *64*(6), 1068-1073. https://doi.org/10.1016/j.jaad.2010.03.039

[84]. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S.,

Nanda, J., & Weber, J. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, *8*. https://doi.org/10.1038/s41597-021-00815-z

[85]. Sadia, Z., & Maishar Mostari, Z. (2022). The Impact of Interactive Dashboard Visualization on Data-Driven Decision Making in Financial Services: A Qualitative Evaluation. *International Journal of Scientific Interdisciplinary Research*, *12*(03), 01–35. https://doi.org/10.63125/3jstsv75

[86]. Sangers, T. E., Wakkee, M., Moolenburgh, F. J., Nijsten, T., & Lugtenberg, M. (2023). Towards successful implementation of artificial intelligence in skin cancer care: A qualitative study exploring the views of dermatologists and general practitioners. *Archives of Dermatological Research*, *315*(5), 1187-1195. https://doi.org/10.1007/s00403-022-02492-3

[87]. Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377-400. https://doi.org/10.1177/0018720816634228

[88]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV),

[89]. Shaiful, M., Anisur, R., & Md, A. (2022). A Systematic Literature Review on the Role of Digital Health Twins in Preventive Healthcare For Personal and Corporate Wellbeing. *American Journal of Interdisciplinary Studies*, *3*(04), 1-31. https://doi.org/10.63125/negjw373

[90]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 432–477. https://doi.org/10.63125/q3r4kk16

[91]. Stabak, D., & Md Siam, T. (2023). Real-Time Edge Computing for Cyber-Physical Systems: A Meta-Analysis of Energy Efficiency and Fault Tolerance. *American Journal of Scholarly Research and Innovation*, *13*(06), 01–44. https://doi.org/10.63125/g9c9cp41

[92]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. *Review of Applied Science and Technology*, *3*(04), 448–508. https://doi.org/10.63125/ahvas976

[93]. Syed Feroz, A. (2023). Quality-Management Systems for Resilient Energy Infrastructure Fabrication. *International Journal of Scientific Interdisciplinary Research*, *13*(10), 37–91. https://doi.org/10.63125/mbhhrm88

[94]. Syed Feroz, A. (2024). AI-Assisted NDT Analytics for Early Crack Detection in Energy-Infrastructure Pressure Vessels Using Advanced Manufacturing Data. *Review of Applied Science and Technology*, *14*(07), 01-52. https://doi.org/10.63125/8bdqs538

[95]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. *International Journal of Scientific Interdisciplinary Research*, *12*(10), 01–53. https://doi.org/10.63125/70msp878

[96]. Tahmedur, R., & Rifat, C. (2023). Critical Path Analysis and Resource-Leveling Optimization for Multi-Project Engineering Portfolios. *American Journal of Interdisciplinary Studies*, *13*(12), 01-47. https://doi.org/10.63125/8eqnea46

[97]. Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., & Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, *20*(7), 938-947. https://doi.org/10.1016/s1470-2045(19)30333-x

[98]. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human-computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229-1234. https://doi.org/10.1038/s41591-020-0942-0

[99]. 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*. https://doi.org/10.1038/sdata.2018.161

[100]. 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

[101]. Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., & Haenssle, H. A. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, *155*(10), 1135-1141. https://doi.org/10.1001/jamadermatol.2019.1735

[102]. Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*. https://doi.org/10.1016/j.artint.2022.103839

[103]. Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P.-A. (2017). Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, *36*(4), 994-1004. https://doi.org/10.1109/tmi.2016.2642839