Enhancing skin lesion classification: a CNN approach with human baseline comparison

Deep Ajabani[1,*], Zaffar Ahmed Shaikh[2,3,*], Amr Yousef[4,5], Karar Ali[6] and Marwan A. Albahar[7]

[1] Source InfoTech Inc., Loganville, Georgia, United States
[2] Department of Computer Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi, Pakistan
[3] School of Engineering, École Polytechnique Federale de Lausanne, Lausanne, Switzerland
[4] Electrical Engineering Department, University of Business and Technology, Jeddah, Saudi Arabia
[5] Engineering Mathematics Department, Alexandria University, Alexandria, Egypt
[6] VentureDive Pvt. Limited, Karachi, Pakistan
[7] College of Engineering and Computing in Al-Lith, Umm Al-Qura University, Makkah, Saudi Arabia
[*] These authors contributed equally to this work.

ABSTRACT

This study presents an augmented hybrid approach for improving the diagnosis of malignant skin lesions by combining convolutional neural network (CNN) predictions with selective human interventions based on prediction confidence. The algorithm retains high-confidence CNN predictions while replacing low-confidence outputs with expert human assessments to enhance diagnostic accuracy. A CNN model utilizing the EfficientNetB3 backbone is trained on datasets from the ISIC-2019 and ISIC-2020 SIIM-ISIC melanoma classification challenges and evaluated on a 150-image test set. The model's predictions are compared against assessments from 69 experienced medical professionals. Performance is assessed using receiver operating characteristic (ROC) curves and area under curve (AUC) metrics, alongside an analysis of human resource costs. The baseline CNN achieves an AUC of 0.822, slightly below the performance of human experts. However, the augmented hybrid approach improves the true positive rate to 0.782 and reduces the false positive rate to 0.182, delivering better diagnostic performance with minimal human involvement. This approach offers a scalable, resource-efficient solution to address variability in medical image analysis, effectively harnessing the complementary strengths of expert humans and CNNs.

Subjects Bioinformatics, Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Neural Networks

Keywords Convolutional neural networks, EfficientNetB3, International skin imaging collaboration, Skin cancer diagnosis, EfficientNet, Deep learning, Medical imaging, Machine learning, International skin imaging collaboration (ISIC), CNN

INTRODUCTION

Early and precise detection of skin lesions is essential for effective treatment and improved patient outcomes (*Houssein et al., 2024*; *Ali et al., 2022*). Despite advancements in medical imaging technologies, significant challenges remain in achieving high diagnostic accuracy and efficiency in real-world clinical environments (*Jackson et al., 2025*; *Esteva et al., 2019*). Skin cancer stands out as a prevalent and aggressive cancer type, impacting over five

Submitted 27 May 2024
Accepted 11 March 2025
Published 15 April 2025

Corresponding authors
Zaffar Ahmed Shaikh, zashaikh@bbsul.edu.pk
Amr Yousef, a.yousef@ubt.edu.sa

Academic editor
Bilal Alatas

Additional Information and Declarations can be found on page 28

DOI 10.7717/peerj-cs.2795

How to cite this article Ajabani D, Shaikh ZA, Yousef A, Ali K, Albahar MA. 2025. Enhancing skin lesion classification: a CNN approach with human baseline comparison. **PeerJ Comput. Sci. 11:e2795 DOI 10.7717/peerj-cs.2795**

million individuals annually around the world (*Liu et al., 2020*; *Kurvers et al., 2019*). Swift and accurate diagnosis is key to effective treatment, prompting substantial investment in refining diagnostic tools (*Kiziloluk et al., 2024*; *Shaikh, 2009*).

Existing literature on skin lesion classification highlights three primary research categories, each addressing distinct challenges in medical diagnostics (*Ahmad et al., 2024*; *Zalaudek et al., 2006*). The first focuses on enhancing human decision-making processes to bolster accuracy (*Yang et al., 2024*; *Kurvers et al., 2021a*, *2021b*). This involves consolidating expert opinions (*Liu et al., 2024*) or devising diagnostic techniques that yield heightened precision (*Combalia et al., 2019*; *Brinker et al., 2018*). However, while these efforts have improved diagnostic accuracy, they are resource-intensive, requiring multiple experts for consensus (*Hernández-Pérez et al., 2024*; *Kousar et al., 2024*).

The second category revolves around artificial intelligence (AI) (*Hosseinzadeh et al., 2024 Sadeghi et al., 2024*), particularly advancements in convolutional neural networks (CNN) for malignancy classification in medical images (*Chatterjee, Gil & Byun, 2024*; *Shaikh et al., 2022*). Several studies have demonstrated that CNNs outperform individual clinicians in skin cancer detection tasks, underscoring their potential to enhance diagnostic accuracy (*Chen et al., 2024*; *Bingol & Alatas, 2021*; *Rezvantalab, Safigholi & Karimijeshni, 2018*). Yet, the generalizability of AI systems remains a challenge, as most studies are conducted under controlled conditions, making it difficult to translate these results directly into clinical settings (*Haenssle et al., 2020*; *Hekler et al., 2019*; *Haenssle et al., 2018*).

The third category explores hybrid models that integrate the first two categories of human expertise with AI predictions, to achieve superior performance (*Tschandl et al., 2019*; *Brinker et al., 2018*; *Shaikh & Lashari, 2017*). For example, studies have shown that ensemble models combining the opinions of multiple clinicians tend to outperform individual diagnoses, provided that their performance levels are comparable (*Kurvers et al., 2019*). However, when performance varies significantly among the experts, ensemble methods may underperform compared to the top-performing individual clinician (*Chatterjee, Gil & Byun, 2024*). This variability highlights the complexity of human-machine collaboration, where balance and structure are essential for optimal results (*Jafar et al., 2024*; *Shaikh et al., 2022*; *Han et al., 2018*).

Similarly, *Haenssle et al. (2020)*, *Marchetti et al. (2020)*, *Brinker et al. (2019)*, *Kurvers et al. (2019)*, *Mahbod et al. (2019)*, *Haenssle et al. (2018)*, and *Esteva et al. (2017)* used CNNs and compared them against human experts in classifying skin lesions through image analysis. All these studies consistently showcased the superiority of machines over humans in this domain. This assertion supports multiple other studies (*Akram et al., 2025*; *Ali et al., 2022*; *Marchetti et al., 2020*). A comprehensive analysis by *Haggenmüller et al. (2021)* concluded that all 19 studies they covered in their comparison demonstrated superior or at least equivalent performance of CNN-based classifiers compared with clinicians.

Despite these advancements, there remains a critical gap in practical and scalable solutions that combine human and machine intelligence while considering real-world constraints (*Park et al., 2023*; *Brady & Neri, 2020*; *Topol, 2019*). Many of the comparative studies mentioned above were conducted under controlled conditions. Furthermore, as highlighted by *Houssein et al. (2024)*, *Nugroho, Ardiyanto & Nugroho (2023)*, *Cassidy et al. (2022)*, and *Haenssle et al. (2020)*, many comparative studies place clinicians in an unfamiliar setting by requiring them to make predictions solely from images, without access to other clinical information. This study addresses this gap by proposing a hybrid algorithm, coined the "augmented hybrid approach". This approach aims to optimize diagnostic performance through selective collaboration between human experts and AI models. It offers an economically viable solution with the potential to improve outcomes and save lives (*Tao & Alatas, 2024*; *Shaikh et al., 2021a*; *Mahbod et al., 2019*).

This augmented hybrid approach extends the concept outlined in studies by *Pirrera & Giansanti (2023)*, *Brady & Neri (2020)*, *Topol (2019)*, *Han et al. (2018)*, which enhances predictive capabilities by providing clinicians with CNN prediction scores as supplementary information.

In this scenario, when a human expert is uncertain about a diagnosis and the algorithm demonstrates high confidence, the expert defers to CNN's prediction, leading to more informed decision-making. This methodology assumes that both human and algorithmic performances align with their confidence in their respective predictions. However, humans often struggle to accurately estimate their confidence, leading to suboptimal use of AI-generated insights (*Akhund et al., 2024a*; *Ha, Liu & Liu, 2020*; *GitHub, 2024a*). As a result, augmented intelligence may not reach its full potential if CNN predictions are underutilized by humans (*Ali et al., 2023*; *Shaikh, 2018*). To address this, our study proposes an algorithmic framework that reverses the traditional approach of relying on human confidence. We use CNN prediction confidence as a proxy for certainty. By replacing human responses with CNN predictions in cases where the network exhibits high uncertainty, we demonstrate a significant improvement in overall performance. This approach is based on prior research, focusing on CNN prediction certainty rather than human confidence estimation (*Ahmad et al., 2024*; *Deotte, 2020*).

This augmented hybrid approach aims to enhance diagnostic precision and the efficient utilization of human resources. Combining the strengths of CNNs with the expertise of medical professionals offers a cost-effective solution to reduce clinician workload and ultimately elevate the quality of treatment (*De, Mishra & Chang, 2024*; *Saeed et al., 2024*; *Dayananda et al., 2023*; *Secinaro et al., 2021*).

For performance evaluation, this research develops a hybrid algorithm that employs the EfficientNetB3 backbone for CNN training, utilizing the ISIC-2019 and ISIC-2020 datasets (*ISIC, 2024*)—the two comprehensive, widely popular, and open-access datasets used in the SIIM-ISIC (Society for Imaging Informatics in Medicine—International Skin Imaging Collaboration) melanoma classification challenges (*Saghir, Singh & Hasan, 2024*; *ISIC, 2024*; *Tan & Le, 2019*; *Codella et al., 2018*; *Gutman et al., 2016*).

By comparing the performance of this hybrid model with both the baseline CNN and human experts, we demonstrate the feasibility and benefits of integrating AI into dermatological diagnostics of skin lesion classification (*Farea et al., 2024*; *Gholizadeh, Rokni & Babaei, 2024*; *Pirrera & Giansanti, 2023*; *GitHub, 2024a*, *2024b*; *Kassem et al., 2021*; *Kassani & Kassani, 2019*; *Han, Mao & Dally, 2015*). This study suggests a hybrid approach that combines CNN predictions with human expertise, potentially improving diagnostic performance by mitigating human errors and machine inaccuracies.

The ISIC-2019 and ISIC-2020 datasets (*ISIC, 2024*) provide benchmarks for training and evaluation, ensuring alignment with current practices and validated methodologies (*Jackson et al., 2025*; *Gouda et al., 2022*; *Adegun & Viriri, 2021*). These datasets also highlight the effectiveness of EfficientNet models, which are recognized for their computational efficiency (*Tan et al., 2024*; *Debelee, 2023*; *Tan & Le, 2019*). In this study, data augmentation techniques are employed to further enhance the performance and generalization of the EfficientNet model across diverse medical images (*Kumar et al., 2024*; *Batool & Byun, 2023*; *Hekler et al., 2020*). This adaptability is essential for real-world clinical settings, where imaging conditions vary significantly (*Shaikh et al., 2022*; *Shorten & Khoshgoftaar, 2019*). Additionally, we collected expert evaluations from 170 medical professionals, primarily dermatologists, to ensure the reliability and robustness of the comparative analysis. From this pool, 69 participants with extensive experience in dermoscopy were selected based on stringent inclusion criteria, reflecting real-world clinical scenarios and providing meaningful insights into the proposed augmented hybrid model's effectiveness. Ultimately, this work contributes to advancing medical imaging by offering a scalable, efficient, and reliable framework for AI-assisted diagnosis (*Akhund et al., 2024b*; *Alam et al., 2022*; *Khamparia et al., 2021*; *Mahbod et al., 2019*; *Han, Mao & Dally, 2015*).

The article is as follows; related studies are discussed in the following section. The section provides an analysis of data, including preprocessing, model selection, computing infrastructure, data evaluation, training data, and train-test split. In "Empirical Study" we conduct research, which involves choosing the model and assessing the metrics. "Training a CNN" covers the training of CNN detailing our attempt and the challenges faced. Our baseline experiments are outlined in "Baseline Experiments". The hybrid algorithm is explained in "Hybrid Algorithms". We discuss the limitations of our study in "Study Limitations" and conclude the article in "Conclusion" by presenting our study results and suggesting areas for further research.

RELATED WORKS

This section aims to represent the works related to the present objectives of the study. There are three primary research directions.

Aggregation of expert opinions

The first research direction focuses on aggregating opinions, which proves beneficial when the ensemble members exhibit similar performance levels (*Hosseinzadeh et al., 2024*; *Freeman et al., 2020*; *Kay et al., 2018*; *Lane et al., 2017*; *Shaikh & Khoja, 2012*). This approach stands to enhance clinical performance but demands substantial resources, as a majority consensus requires soliciting input from multiple clinicians for each case (*Wubineh, Deriba & Woldeyohannis, 2024*; *Hussain et al., 2023*; *Bejnordi et al., 2017*).

CNNs *vs.* human experts

The second research direction asserts that CNNs demonstrate comparable or superior proficiency to humans in categorizing skin lesion malignancies (*Navarrete-Dechent, Liopyris & Marchetti, 2020*). However, further inquiry remains imperative in this domain as current algorithms have yet to reach a level where they can replace human classification (*Chatterjee, Gil & Byun, 2024*; *Esteva et al., 2017*). Consequently, investigating hybrid methodologies represents a logical progression in research (*Hasan et al., 2022*). Various hybrid approaches have been presented, demonstrating superior performance compared to individual human or machine capabilities (*Bozkurt, 2023*; *Goceri, 2020*). Despite the potential of hybrid intelligence to save lives (*Brinker et al., 2019*; *Liu et al., 2020*; *Tschandl et al., 2019*; *Ketkar & Santana, 2017*), a critical gap in the literature pertains to the expenses associated with such methods (*Wubineh, Deriba & Woldeyohannis, 2024*; *Chollet & Chollet, 2021*), which can make even highly effective approaches impractical for real-world implementation (*Topol, 2019*). For instance, the study by *Hekler et al. (2019)* involved the use of multiple clinicians to assess a single image, which would be infeasible in clinical practice.

Hybrid methodologies and the challenges of AI

Similar findings were observed by *Houssein et al. (2024)*, *Koçak et al. (2024)*, *Secinaro et al. (2021)*, *Tschandl et al. (2020)*, who highlighted that less experienced clinicians benefited the most from computer predictions. However, they cautioned that faulty AI could mislead clinicians across all experience levels, posing a threat to the usability of hybrid approaches (*Kaluarachchi, Reis & Nanayakkara, 2021*; *Han et al., 2018*; *Chollet & Chollet, 2021*). Blindly trusting AI decisions could undermine the superior performance achieved through human-machine collaboration, as demonstrated in these studies (*Brinker et al., 2018*; *Goodfellow, 2016*; *Shaikh & Khoja, 2011*). For example, *Ha, Liu & Liu (2020)*, *Marchetti et al. (2020)* pursued a systematic approach where human participants rated their confidence in each image. By supplementing low-confidence ratings with machine-generated predictions, performance improved, particularly for medical residents, but less so for professional dermatologists. Although this method mitigates the risk of over-reliance on CNN predictions, it remains susceptible to flawed AI (*Gómez-Carmona et al., 2024*; *Krakowski et al., 2024*; *Keerthana, Venugopal & Nath, 2023*).

Ensemble approaches for skin lesion classification

The third research direction is the use of ensemble approaches (*Aboulmira et al., 2025*; *Hekler et al., 2019*). By aggregating predictions using a boosting algorithm, *Gulli & Pal (2017)* found that combining human and computer decisions yielded the best results for multiclass and binary lesion classification. This ensemble approach prevents human reliance solely on machine-generated results and appears to be a more feasible real-world implementation (*Yang et al., 2024*; *Reddy et al., 2024*). However, comprehensive comparisons of these approaches are still lacking (*Ali et al., 2023*). In the examination of existing literature (*De, Mishra & Chang, 2024*; *Kumar et al., 2024*; *Tschandl et al., 2020*; *Shaikh & Khoja, 2014*), the majority of studies focused on skin lesion classification relied on the utilization of one or more pre-trained networks as a foundational framework for their models.

In Table 1 we present a well-rounded comparative analysis of recent advancements, hybrid models, and interpretability-focused essential and recent studies in skin lesion classification literature that align closely with the objectives and key themes of this study.

DATA

This study utilizes three distinct datasets for analysis: the original BCN20000 dataset of the three-point checklist of dermoscopy (*Combalia et al., 2019*) and the ISIC-2019 and ISIC-2020 datasets (*ISIC, 2024*). The BCN20000 dataset comprises 165 images alongside corresponding human responses. This dataset is deemed suitable for constructing and validating the ensemble technique (*Freeman et al., 2020*; *Shaikh et al., 2019*), serving as the evaluation dataset herein. However, its size proves insufficient for adequate CNN training (*Saeed et al., 2024*; *Batool & Byun, 2023*; *Sun et al., 2023*). Therefore, substantially larger datasets, ISIC-2019 and ISIC-2020, denoted as the training datasets, were employed for this purpose. The subsequent sections delineate the procedures involved in gathering and preprocessing these datasets, with priority given to an initial overview of the evaluation dataset.

Data preprocessing

During the research, preparing the data was crucial to train CNN models efficiently (*Pérez & Ventura, 2022*; *Marchetti et al., 2020*). These initial steps were necessary to get the dataset ready, for training and evaluation guaranteeing that the CNNs could learn well and provide predictions (*Koçak et al., 2024*; *Hekler et al., 2020*; *Brinker et al., 2019*).

All images were uniformly downsized to a resolution of 256 × 384 pixels. This decision balanced the need for detail preservation and computational efficiency, as higher resolutions would increase training time without guaranteeing better performance (*Aboulmira et al., 2025*; *Ali et al., 2021*). To enhance the training dataset, an ImageDataGenerator was used to perform various augmentation techniques such as rotation, zoom, and horizontal flipping (*Bozkurt, 2023*; *Shaikh et al., 2021b*; *Goceri, 2020*). This approach helped in creating a more robust model by exposing it to a variety of image transformations, thereby improving generalization (*Hekler et al., 2020*; *Han et al., 2018*).

During the training and validation phases, we opted for a batch size of 64 to ensure a mix of classes, in each batch which is important for addressing class imbalances (*Aboulmira et al., 2025*). When it came to testing, we used a batch size of 1 to maintain the image order and ensure alignment between predictions and actual data (*Shaikh et al., 2024*; *Shaikh & Khoja, 2014*). The images were normalized to standardize the input data aiding in speeding up the CNNs learning process by ensuring that the data distribution has an average of zero and a standard deviation of one (*Chollet & Chollet, 2021*; *Brinker et al., 2019*).

Table 1 **Summary of pre-trained models.**

Reference	Network(s) used
Brinker et al. (2019)	ResNet50
Abdelrahman & Viriri (2023)	EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7, Se-ResNext101, ResNest101
Han et al. (2018)	ResNet-152
Hekler et al. (2020)	ResNet50
Haenssle et al. (2018)	InceptionV4
Haenssle et al. (2020)	MoleAnalyzerPro
Li et al. (2020)	Custom
Esteva et al. (2017)	Inception V3
Haggenmüller et al. (2021)	AlexNet, VGG16, VGG19, GoogleNet, ResNet-50, ResNet-101, ResNet-152, Inception-V3, Inception-V4, DenseNets, SeNets, PolyNets
Han et al. (2020)	SeNet, Se-ResNet50, VGG19
Tschandl et al. (2020)	ResNet34

Justification of model types and selection method

Model types

This study employed EfficientNetB3 models for skin lesion classification, chosen for their balance of high performance and computational efficiency (*Aboulmira et al., 2025*; *Kassem et al., 2021*; *Tan & Le, 2019*). The justification for selecting EfficientNetB3 and the model types used are based on several key factors.

Performance and efficiency

EfficientNets have shown better results than networks having a similar parameter count, as seen in *Hasan et al. (2021)*, *Huang et al. (2022)*. They deliver outcomes in tasks, like image classification while being computationally effective making them a budget-friendly option for this research (*Aboulmira et al., 2025*; *Ali et al., 2022*; *Zalaudek et al., 2006*). EfficientNet model EfficientNetB3 uses parameters compared to various conventional CNN designs (*Alhichri et al., 2021*; *Li et al., 2018*). This efficiency helps decrease the workload during training, which is essential considering the limited computing resources accessible for this study (*Akhund et al., 2024a*; *Ha, Liu & Liu, 2020*; *Hekler et al., 2020*).

Demonstrated effectiveness

EfficientNets, especially the EfficientNetB3 variant, were prominently used in high-ranking submissions of the 2019 and 2020 SIIM-ISIC melanoma classification challenges (*ISIC, 2024*). This track record of success in similar dermatological imaging tasks is also seen in *Reddy et al. (2024)*, *Gouda et al. (2022)*, and *Feng et al. (2022)*, underscores their suitability for skin lesion analysis.

To prevent overfitting, the study employed various regularization techniques, including data augmentation and dropout layers (*Marchetti et al., 2020*; *Srivastava et al., 2014*; *Argenziano et al., 2003*). EfficientNetB3's design allows for the integration of these techniques (*Aboulmira et al., 2025*; *Kim & Bae, 2020*), further enhancing its performance on the validation set and ensuring better generalization to new data (*Salman & Liu, 2019*).

Selection method

The selection method for determining the most suitable model involved the study exploring a grid of models with variations in label encoding (binary/multiclass), model head capacity (shallow/deep), and dropout layer strength (0/0.2/0.4/0.6) (*Jackson et al., 2025*; *Bergstra & Bengio, 2012*). This comprehensive grid search identified the optimal combination of these parameters to maximize model performance (*Tuba et al., 2021*). Early stopping was implemented to terminate training when no improvement in validation loss was observed for seven consecutive epochs (*Tuba et al., 2021*; *Yu, Song & Ren, 2013*). This approach ensured that models did not overfit, and training resources were used efficiently (*Saghir, Singh & Hasan, 2024*; *LeCun et al., 1998*; *Smith, 2017*; *Dietterich, 1995*).

To expedite the training process, models were trained in parallel and grouped by target label and model capacity (*Nugroho, Ardiyanto & Nugroho, 2023*; *Huang et al., 2017*). This strategy significantly reduced the total training time, allowing for a thorough exploration of the model grid within a feasible timeframe (*Akram et al., 2025*; *Brinker et al., 2019*). After optimizing models on the validation set, their performance was evaluated on an independent test set (*Huang et al., 2022*; *Abadi et al., 2016*). This final evaluation step ensured that the selected models generalize well to new, unseen data (*Loshchilov & Hutter, 2016*).

In summary, EfficientNetB3 was chosen for its demonstrated performance and efficiency (*Chollet & Chollet, 2021*; *Tschandl et al., 2019*), with the selection method ensuring robust and generalizable model performance through a systematic and resource-efficient training process (*Hosseinzadeh et al., 2024*).

Computing infrastructure

We required enormous computing power to conduct this research because of the nature and large scale of the deep learning models and datasets utilized (*NVIDIA Corporation, 2020*).

The research made use of Linux-based operating systems for their reliability and efficiency (*Gulshan et al., 2016*), in handling tasks and compatibility with various deep learning frameworks (*Khalil et al., 2023*). Training learning models, such as CNNs like EfficientNet demand GPU power (*Jouppi et al., 2017*). The study employed high-performance NVIDIA GPUs optimized for learning tasks (*Kang & Tian, 2018*; *NVIDIA Corporation, 2020*). While the research considered solutions utilizing TPUs (Tensor Processing Units) known for their efficiency in tensor operations in networks, the practical implementation primarily depended on GPU resources due to their availability and infrastructure limitations (*Jouppi et al., 2017*; *Abadi et al., 2016*; *Nair & Hinton, 2010*). Multi-core CPUs played a role in preprocessing data and overseeing workflow management.

Handling large datasets such as BCN20000, ISIC-2019, and ISIC-2020 requires high-capacity storage solutions (*Ali et al., 2022*; *Nguyen et al., 2019*; *ISIC, 2024*). Fast SSDs were used to ensure quick data access and processing speeds (*Aboulmira et al., 2025*; *Kumar et al., 2024*). Training complex models and processing large datasets necessitated large amounts of RAM to handle data efficiently during training phases.

The main tool for developing and training models was TensorFlow (*Jang, 2025*; *Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*; *Abadi et al., 2016*). This framework is widely supported on Linux (*Abadi et al., 2016*; *Nair & Hinton, 2010*). It provides a range of tools for creating and improving neural networks. Kaggle Notebooks were used for model testing (*Mostafavi Ghahfarokhi et al., 2024*; *Mukhlif et al., 2024*). To benefit from community-shared solutions, these notebooks offered an adaptable environment for conducting experiments with computing requirements. Various personalized scripts—for processing, enhancing, and assessing the data—were created to customize the workflows according to the study's needs (*Akhund et al., 2024b*; *Banachewicz & Massaron, 2022*; *Abadi et al., 2016*).

The computing setup relied on GPUs, ample memory, spacious storage, and reliable deep-learning software, on a Linux system to support smooth and productive model training and assessment procedures (*Haggenmüller et al., 2021*; *Nair & Hinton, 2010*).

Evaluation data

Image dataset description

The BCN20000 dataset comprised 165 images selected randomly from a larger collection of 2,621 images. The sole criteria for inclusion were adequate image quality and the presence of hemoglobin pigmentation in either the entire lesion or part thereof (*Mostafavi Ghahfarokhi et al., 2024*; *Combalia et al., 2019*; *Esteva et al., 2017*). Among these 165 images, 15 were allocated for training purposes, facilitating participant familiarity with the process of utilizing the three-point checklist for evaluation. The remaining subset of 150 images served as the basis for assessing performance in this study. This set of 150 images, along with the corresponding human evaluations, forms the core of the evaluation dataset employed in this research endeavor (*Rotemberg et al., 2021*; *Tschandl et al., 2020*). Specifically, the BCN20000 dataset offered the following components:

- JPEG (JPG) files with a resolution of 512 × 768 pixels for each image.
- Individual evaluations by participants utilizing the three-point checklist for each image.
- Ground truth data corresponding to each image.
- Metadata associated with the images.
- Metadata associated with the participating individuals

Each image in the BCN20000 dataset was presented at a resolution of 512 pixels in height and 768 pixels in width, thereby establishing an aspect ratio of 1:1.5 (*Hernández-Pérez et al., 2024*). To ensure consistency in training and testing processes with similar images, the training images referred to in the preceding section were adjusted to the same aspect ratio (*Tan et al., 2024*; *Combalia et al., 2019*).

Image classification and ground truth

The investigation focuses on categorizing images through binary classification, distinguishing between a "benign" and a "malignant" category. The benign class represents a negative status denoting no cause for concern, while the malignant class signifies the presence of cancer (*Chatterjee, Gil & Byun, 2024*; *Marchetti et al., 2020*). Study participants were not directed to categorize images as benign or malignant; instead, they assessed each image based on three distinct characteristics: asymmetry (about color and/or structure, not shape), atypical network (characterized by a pigment network displaying thick lines and irregular holes), and blue-white structures (indicating the presence of blue and/or white coloration within the lesion) (*Mostafavi Ghahfarokhi et al., 2024*; *Tschandl et al., 2019*). When two or more of these characteristics were identified, the lesion was classified as malignant. To establish malignancy scores for each participant and image, the responses for each criterion were converted into binary form, where 0 represented "not present" and 1 indicated "present". The analysis discerns the classification of images into negative or positive classes for each participant (*Jackson et al., 2025*; *Marchetti et al., 2020*). The ground truth of each image was established *via* histopathological examination (*Esteva et al., 2019*). Among the 165 images, 116 were benign instances and 49 were malignant instances, indicating an incidence rate of 29.7% (*Haenssle et al., 2018*). This rate notably exceeds that of the training images, which could potentially impact the outcomes (*Marchetti et al., 2020*). Accompanying each image is metadata detailing the subject of the image, encompassing age, sex, and the lesion's anatomical location, akin to the information provided in the ISIC-2019 and ISIC-2020 datasets (*ISIC, 2024*; *Tan & Le, 2019*). Nonetheless, the decision was made to exclude metadata from the model.

Participant selection criteria

The evaluation encompassed 170 participants, who provided background information related to their professional roles and medical experience. This information included details such as:

- **Professional background:** Participants were primarily dermatologists or medical professionals with substantial expertise in skin lesion analysis.
- **Country of origin:** The participants came from various regions, ensuring geographic diversity. This may have helped capture a range of diagnostic approaches and perspectives.
- **Experience with dermoscopy:** Participants were asked to report their prior experience with dermoscopy, including whether they were routinely engaged in diagnosing skin lesions through dermoscopy in clinical practice.
- **Years of experience:** The number of years participants had been performing dermoscopies was recorded to ensure that only individuals with sufficient experience were included in the performance evaluation.
- **Frequency of yearly dermoscopies:** To further quantify their expertise, participants also reported how often they perform dermoscopies each year. This helped in distinguishing between frequent and less frequent users of the technique.

For the scope of this study, the inclusion was limited to participants who had completed at least 126 out of the 150 images in the evaluation dataset. This threshold was set to ensure a reliable assessment of each participant's performance. Based on this criterion, 69 experienced participants were selected for the analysis. This approach was designed to reflect a real-world clinical scenario and to provide a meaningful comparison between human evaluators and the CNN model (*Lee et al., 2025*).

Training data

A substantial volume of high-quality images is crucial to train a CNN effectively (*Ali et al., 2022*; *Rotemberg et al., 2021*). For instance, the ImageNet Challenge used 1.2 million images across 1,000 categories, averaging 1,200 images per category (*Hekler et al., 2019*; *Gutman et al., 2016*; *Deng et al., 2009*). However, gathering and verifying these images is labor-intensive, resulting in limited datasets for computer vision tasks (*Tan et al., 2024*). Exploring skin lesion datasets, the PH2 dataset provided only 200 images, insufficient for CNN training (*Gouda et al., 2022*). The SIIM-ISIC challenges between 2016 and 2020 saw a significant expansion (*ISIC, 2024*) as shown in Table 2.

The ISIC-2020 dataset included 33,126 images (*ISIC, 2024*; *Rotemberg et al., 2021*), curated from various sources after thorough quality checks. Despite the volume, the ISIC-2020 dataset had a low positive incidence rate of 1.76%, potentially causing imbalances in training. To address this, merging the BCN20000 dataset with a 17.85% incidence rate was necessary, providing more positive instances for better model learning. Both datasets were used for this reason. The inclusion of metadata (patient age, sex, lesion site) in skin lesion classification models has shown promise. Studies suggest that incorporating such data enhances diagnostic accuracy for clinicians (*Reddy et al., 2024*; *Haggenmüller et al., 2021*). However, some winning Kaggle models (*Lee et al., 2025*; *Ha, Liu & Liu, 2020*; *Lopez et al., 2017*) did not benefit from metadata fusion, opting for different strategies. *Haenssle et al. (2018)* also highlight the potential of metadata but suggest varying benefits depending on the model structure (*Naseri & Safaei, 2025*; *Codella et al., 2018*). Aligning labels for both datasets, as shown in Table 3, involved treating the "MEL" class as positive and others as negative. Yet, using multiclass labels poses a risk: it might improve performance but limit overall utility (*Shen et al., 2019*; *Deng et al., 2009*). To counter this, we experimented with training models using both binary and multiclass labels.

Train-test split

Neural networks, due to their high capacity, are prone to overfitting, making it unsuitable to evaluate them on the same data used for training (*Abbas et al., 2025*; *Ophir et al., 1991*). Standard practice involves splitting the data into three sets: training, validation, and test sets. The training set is used to optimize the model's performance, but overfitting can occur if the model memorizes the training data (*Alotaibi & AlSaeed, 2025*; *Salman & Liu, 2019*). A validation set, derived from the training set, helps identify overfitting by monitoring the loss on both training and validation sets after each epoch (*Liu et al., 2025*; *LeCun et al., 1998*). Overfitting is detected when training loss decreases while validation loss increases (*Srivastava et al., 2014*). The validation set also determines when to stop training to avoid further overfitting (*Debelee, 2023*). However, models optimized excessively on the validation set risk "overfitting to the validation set", where some models perform better by chance rather than reflecting the true model. To mitigate this, a test set—a separate subset—is used to assess the final model's ability to generalize new, unseen data. After

Table 2 ISIC competition in the year 2016–2020.

Year	# of images
2016	900
2017	2,000
2018	12,609
2019	25,331
2020	33,126

Table 3 Alignment procedure for the 2019 and 2020 ISIC data labels.

2019 Diagnosis	2020 Diagnosis	Target
NV	Nevus	NV
MEL	Melanoma	MEL
BCC	BCC	BCC
BKL	Seborrheic keratosis, lichenoid keratosis, solar lentigo, lentigo NOS	BKL
AK		AK
SCC		SCC
VASC		VASC
DF		DF
	Cafe-au-lait macule, atypical melanocytic proliferation, unknown	Unknown

optimizing the validation set, the model's performance on the test set ensures its suitability for predicting novel images (*Aboulmira et al., 2025*; *Tan & Le, 2019*).

In this study, we employed a 15% validation split (*Liu et al., 2024*, *2020*; *Chollet & Chollet, 2021*). Due to differing incidence rates in the ISIC-2019 and ISIC-2020 datasets (*Hernández-Pérez et al., 2024*; *ISIC, 2024*; *Rotemberg et al., 2021*), a stratified train-validation split was implemented, using an 85-15 division for each dataset before combining them (*Naseri & Safaei, 2025*; *Ali et al., 2022*; *Deotte, 2020*). Table 4 shows the class distribution for these splits. The test set provided by the 2020 Kaggle competition was used directly, omitting the need for a custom test set. This test set of the ISIC-2020 dataset comprises 10,982 images without ground truth, preventing model-specific tuning (*ISIC, 2024*). Instead, Kaggle accepts model predictions for scoring, facilitating performance comparisons with competition participants (*ISIC, 2024*; *Ha, Liu & Liu, 2020*).

EMPIRICAL STUDY

An analysis was conducted with more than 30 pre-trained networks available for selection. Table 1 comprehensively illustrates the prevalent use of ResNet networks as foundational structures for skin lesion analysis. However, the prominent submission in the 2020 SIIM-ISIC melanoma classification challenge predominantly leveraged EfficientNets to highlight the consistently superior performance of EfficientNets compared to other networks with similar parameter counts (*ISIC, 2024*; *Tan et al., 2024*). This suggests these networks might be more cost-effective due to their lower parameter count, demanding less

Table 4 Class distribution in training and validation sets.

Category	Training share %	Validation share %
AK	1.5	1.6
BCC	5.6	6.0
BKL	4.9	4.6
DF	0.4	0.4
MEL	8.8	8.6
NV	31.0	30.3
SCC	1.1	1.1
Unknown	46.3	47.1
VASC	0.4	0.4

computational resources during training (*Gouda et al., 2022*). A comprehensive review of pre-trained networks within the Keras package further substantiates the effectiveness of EfficientNets (*ISIC, 2024*). They exhibit creditable performance in the ImageNet Challenge while necessitating fewer parameters and reasonable training durations (*Houssein et al., 2024*; *Hekler et al., 2019*). Given constrained computational resources, opting for an EfficientNet seems a judicious choice for this study (*Ali et al., 2022*). Specifically, selecting the EfficientNet B3 variant aligns with the models utilized (*Hosseinzadeh et al., 2024*; *Tan & Le, 2019*; *Ophir et al., 1991*).

Regularization

Deep neural networks, specifically Deep CNNs, exhibit extensive model capacity, making them susceptible to overfitting. Techniques employed to counter overfitting are known as regularization methods and encompass various approaches.

In this study, four key regularization techniques—data augmentation, dropout layers, capacity regulation, and weight regularization—are optimized following prior research (*Srivastava et al., 2014*; *Argenziano et al., 2003*). *Chollet & Chollet (2021)* suggests exploring model capacity until overfitting appears, followed by applying regularization methods to improve test performance. This process is iterative, time-consuming, and requires expertise from the data scientist (*Hekler et al., 2019*). Data augmentation prevents overfitting by altering input data, ensuring the model learns general features rather than specific images. This technique is crucial for small datasets and also improves generalization in larger datasets (*Goodfellow, 2016*). Rotation, zooming, and horizontal flipping are used in this study to introduce variations, helping the model handle novel orientations (*Hekler et al., 2019*; *Zalaudek et al., 2010*). Although shear is a common augmentation strategy, *Zalaudek et al. (2006)* highlight its potential to distort asymmetrical features, crucial for detecting malignant melanomas, which led to its exclusion. Based on prior literature observations (*Hernández-Pérez et al., 2024*; *Chollet & Chollet, 2021*; *Goodfellow, 2016*; *Srivastava et al., 2014*), data augmentation remains a vital part of regularization in all models.

Dropout layers, situated between hidden layers in a neural network, employ a binary mask to deactivate part of the activation, compelling subsequent layers to operate with incomplete information (*Hernández-Pérez et al., 2024*; *Srivastava et al., 2014*). This technique effectively converts the model into a correlated ensemble without multiple model instances (*Hekler et al., 2019*). Regulating the capacity of a neural network involves adjusting its depth and width, impacting the number of parameters. EfficientNets were developed by striking a balance between these dimensions, aiming for an optimal structure (*Mukhlif et al., 2024*; *Haggenmüller et al., 2021*; *Tan & Le, 2019*; *Dillon et al., 2017*; *Goodfellow, 2016*). *Chollet & Chollet (2021)* suggest an approach involving the deliberate construction of a complex network, followed by applying regularization techniques to counter overfitting. However, due to the substantial complexity of datasets and models, an alternative strategy of concurrently training multiple models with varied capacities and regularization methods is employed in this study. Weight decay, a regularization technique imposing a penalty function on complexity, reduces the network's tendency to overfit (*Houssein et al., 2024*; *Loshchilov & Hutter, 2016*). L^2 weight decay, a common form, adjusts the relative contribution of the norm penalty function through a weight parameter. Finding an optimal alpha value (determining the degree of regularization) necessitates experimentation tailored to the specific situation (*Sterkenburg, 2025*; *Hastie, Tibshirani & Friedman, 2009*). Initial attempts to implement L^2 weight decay into the network using specific values encountered technical challenges, causing conflicts within TensorFlow and subsequent crashes (*Jang, 2025*; *Pang, Nijkamp & Wu, 2020*; *Hekler et al., 2019*; *Hastie, Tibshirani & Friedman, 2009*). As a result, prioritizing other aspects of the study overcame these obstacles.

Metrics

When assessing the performance of a model in a classification task, a frequently employed and straightforward metric is prediction accuracy (*Mohammed & Meira, 2020*; *Davis & Goadrich, 2006*). This metric, as defined by *Mohammed & Meira (2020)*, serves as a common evaluation criterion:

$$Accuracy \ = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \widehat{y}_i). \tag{1}$$

In a scenario with 'n' observations, accuracy measures how well a model's estimates match the actual answers. Yet, accuracy can fall short for imbalanced datasets, where it may inflate due to a majority class bias, rendering them less useful (*Liaw et al., 2025*; *He & Garcia, 2009*). To counter this, the receiver operating characteristic curve (ROC)/area under curve (AUC) analysis proves valuable, unaffected by such imbalances (*Ozel et al., 2025*; *Ha, Liu & Liu, 2020*; *Esteva et al., 2019*; *Fawcett, 2006*). Additionally, the connection between CNN outputs and ROC thresholds has made ROC analysis prevalent in evaluating skin lesion CNNs. ROC analysis hinges on the true positive rate (TPR) and false positive rate (FPR), explained through confusion matrices (*Gouda et al., 2022*; *Yap, Yolland & Tschandl, 2018*). In a binary classification scenario with 'n' observations, 'y_i' represents true labels, and '$\widehat{y}_i$' signifies estimated labels for each observation indexed from 1 to 'n', distinguishing positive ('c_1') and negative ('c_2') classes (*Hernández-Pérez et al., 2024*; *Hekler et al., 2019*; *Kassem et al., 2021*).

Hence, in line with commonly used metrics in skin lesion classification research, as demonstrated in *Houssein et al. (2024)*, *Ali et al. (2022)*, *Yap, Yolland & Tschandl (2018)*, *Esteva et al. (2019)*, we employ AUC-ROC, TPR, and FPR for performance evaluation.

TRAINING A CNN

This section outlines the prevalent strategy of employing a pre-trained network derived from the ImageNet challenge as the foundation for training a CNN aimed at skin lesion classification (*Reddy et al., 2023*; *Ha, Liu & Liu, 2020*). This method finds frequent application in research articles (refer to Table 1) and has been proven instrumental in securing victory in a significant Kaggle competition (*Banachewicz & Massaron, 2022*). It is endorsed in dedicated deep-learning literature (*Hernández-Pérez et al., 2024*; *Ali et al., 2022*). Subsequent sections delineate endeavors in constructing and training CNN.

Preliminary training approach

The primary objective was to train a CNN using the EfficientNetB3 architecture (*Alhichri et al., 2021*), with a Flattening layer followed by three Dense layers. The goal was to explore a large hyperparameter space that extended beyond the scope outlined in "Empirical study". A full search, excluding learning rate tuning, would require training 1,024 models, which was impractical. To address this, the Keras tuner package Hyperband (*Li et al., 2018*) was used, applying a multi-armed bandit strategy to systematically evaluate models within the hyperparameter space (*Ali et al., 2022*). Hyperband's heuristic approach creates a subset of models, trains them briefly for a few epochs, saves the models, and discards inferior performers (*Alhichri et al., 2021*; *Hekler et al., 2019*). The top-performing models are iteratively refined through further training, allowing resources to focus on the most promising candidates (*Chen et al., 2024*; *Li et al., 2018*). A preliminary test with Hyperband used 320 training images (less than 1% of the total 50,000) and 160 validation images (192 × 158 pixels). Training for 15 epochs took approximately 7 h, during which 90 models were evaluated, identifying the best-performing one. This model's base was unfrozen and fine-tuned with 10 additional epochs. As shown in Table 5, the final model had more parameters in the Dense layers, driven by the Flattening layer's large output size.

However, four issues emerged during the evaluation:

1) **Image dimension error:** Images were incorrectly encoded as (192,158) instead of (128,192), affecting performance (*Goodfellow, 2016*).
2) **Runtime limitations:** Testing on a cloud setup led to suboptimal node utilization, with training marked by performance spikes and inefficiencies. Enlarging the images to (512,768) increased the epoch runtime by 12×, rendering a full-scale test infeasible (*Chollet & Chollet, 2021*).

Table 5 **Model summary of the best model.**

Layer	Output shape	Number of parameters
EfficientNetB3 (functional)	(None, 6, 5, 1,536)	10,783,535
Flatten	(None, 46,080)	0
Dropout	(None, 46,080)	0
Dense1	(None, 256)	11,796,736
BatchNorm	(None, 256)	1,024
Dense2 (Dense)	(None, 160)	41,120
Dense (Dense)	(None, 8)	1,288

3) **Dataset inconsistency:** Initially, only the HAM10000 subset of the 2019 dataset was downloaded, encoding eight classes instead of nine. The complete dataset was later obtained and re-encoded to meet the required specifications (*Tschandl et al., 2019*).

4) **Prediction imbalance:** Despite achieving low loss values, the model's predictions overestimated prevalent classes while assigning low probabilities to less common ones, limiting its generalization (*Bria, Marrocco & Tortorella, 2020*).

Optimizing the training strategy

The initial strategy faced obstacles that required significant time and effort to resolve. An analysis of the winning solution from the 2019 and 2020 Kaggle competition (*GitHub, 2024a*, *2024b*; *Lin et al., 2017*) revealed available code on GitHub (*GitHub, 2024a*, *2024b*), but the training process demanded extensive computational resources, exceeding our infrastructure. Additional exploration uncovered contributions from a Kaggle grandmaster, who utilized TPUs—resources not available to us (*Banachewicz & Massaron, 2022*). Another submission by *Jang (2025)* provided valuable insights but also relied on TPUs, requiring image rescaling to a 1:1 ratio. We evaluated this model after resizing images to (256,256), achieving an AUC score of 0.773. Through iterative reviews and small-scale testing, we identified promising adjustments that improved training by focusing on image features rather than class distributions. Key observations included:

- Implementation of GlobalAveragePooling instead of Flattening layers between the base and head, substantially enhancing performance while reducing the parameters in the initial dense layer (*Ali et al., 2022*; *Minderer et al., 2022*).
- Experimentation with binary-labeled training sets, inspired by Tensorflow (*Jang, 2025*; *Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*), demonstrated quick and satisfactory results despite potential information loss compared to multilabel training (*Houssein et al., 2024*; *Bria, Marrocco & Tortorella, 2020*).
- Adoption of a learning rate schedule aligned with *Chollet & Chollet*'s *(2021)* recommendation, diverging from the reviewed models' approaches, showcased substantial performance enhancements (*Ha, Liu & Liu, 2020*; *Loshchilov & Hutter, 2016*).

- Construction of models with shallow heads akin to Tensorflow (*Jang, 2025*; *Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*) designs resulted in significantly fewer parameters than conventionally proposed by *Chollet & Chollet (2021)*.
- Integration of class weights into the optimizer, following the Tensorflow approach (*Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*), counteracted skewed class distributions, notably improving model predictions (*Lin et al., 2017*).

The collective impact of these methodologies substantially enhanced our models' capability to discern image features rather than relying solely on class distribution information (*Hernández-Pérez et al., 2024*). Consolidating these strategies, a new model is crafted, expected to surpass the evaluation set performance achieved by the Tensorflow model (*Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*).

Final model training and evaluation

The final configuration focused on three key parameters: label encoding (binary/multiclass), model head capacity (shallow/deep), and dropout layer strength (0/0.2/0.4/0.6). Initially, we attempted to use two dropout sizes (0/0.5) and an L2 kernel regularizer in the first Dense layer. However, this setup caused compatibility issues with EfficientNetB3, leading to recurring errors. As L2 regularization was absent in similar models, we shifted focus to exploring more dropout configurations. A total of 16 models were trained with varying parameter combinations.

To balance convergence and training time, we set the maximum epoch limit to 50 with an early stopping callback (patience = 7). The best model, based on validation loss, was retained. Early stopping minimized computational overhead by halting training when no improvement was observed over seven consecutive epochs (*Caruana & Niculescu-Mizil, 2006*).

While EfficientNet models can handle higher resolutions, they increase computational demands without guaranteed performance improvements (*Tan & Le, 2019*). Thus, we downscaled all images to (256,384) for training, validation, testing, and evaluation. Initial trials showed each epoch took just over two hours. Following prior studies (*Ali et al., 2022*; *Shaikh & Khoja, 2013*), we estimated 35–40 h of total training time per model over 15 epochs.

To speed up training, we ran models in parallel, grouped by target label and model capacity. Four separate scripts were developed, each using one of the four dropout configurations. These scripts applied identical preprocessing steps with an ImageDataGenerator to resize and augment images (*Chollet & Chollet, 2021*). Training batches were set to 64 to ensure the representation of smaller categories, while test batches used a size of 1 to maintain image order alignment with predictions. Each script's total execution time ranged between 8 and 12 days.

BASELINE EXPERIMENTS

This section outlines three key aspects. Initially, the performance of 69 relevant individuals on the evaluation dataset is detailed, and their average performance establishes a human baseline. Subsequently, the selection process for determining the CNN baseline among 16 trained CNNs is presented. Finally, an evaluation and comparison of this CNN baseline against both human performances and the Tensorflow (*Pang, Nijkamp & Wu, 2020*; *Dillon et al., 2017*) model is provided.

The performance of each individual is quantified as a pair of FPR and TPR values. The FPR is calculated as the ratio of false positives to the total number of actual negatives, while the TPR, also known as Sensitivity or Recall, is the ratio of true positives to the total number of actual positives (*Hanley & McNeil, 1982*). These metrics are essential for evaluating model performance, especially in medical imaging contexts where class imbalances can significantly impact results.

Establishing a human baseline is critical for understanding the performance of automated systems in clinical settings. Previous studies in dermatology have emphasized the variability of human diagnostic capabilities, often indicating that dermatologists can achieve TPRs in the range of 0.70 to 0.85 (*Rawat, Rajendran & Sikarwar, 2025*; Khan et al., *2024*). The selection process for identifying the CNN baseline among the 16 trained models should involve comparing metrics like AUC-ROC and cross-validation performance to ensure robustness (*Rawat, Rajendran & Sikarwar, 2025*; *Ophir et al., 1991*). Comparing the CNN baseline to human performance allows for assessing the effectiveness of the model in mimicking or surpassing human diagnostic capabilities (*Ali et al., 2022*; *Esteva et al., 2017*).

Figure 1 illustrates the performance of individual humans, accompanied by an average FPR and TPR denoted by a blue dot. This average value does not signify an ensemble value but represents the mean performance level of humans. It serves as a benchmark for human performance for subsequent analysis and comparison. The average FPR stands at 0.196, while the average TPR is 0.765 (*Grzybowski, Jin & Wu, 2024*).

All 16 trained models, as expounded, undergo evaluation using the ROC/AUC metrics stipulated in "Empirical Study" across training, validation, test, and evaluation datasets. The AUC scores for these models are tabulated in Table 6, with the highest scores per category highlighted in boldface. An examination of the table reveals pertinent observations:

1) **Overfitting indicators:** Scores typically peak on the training data, a common occurrence in machine learning owing to model feature acquisition from the training data. Substantial disparities between training and validation/test scores could signal overfitting (*Khan et al., 2025*; *Goodfellow, 2016*). Test scores tend to be marginally lower than validation scores. Two plausible explanations arise. Firstly, dissimilar incidence rates between the training set (derived from the 2019 Kaggle competition) and the test set could lead to biased performance if the model learned this incidence distribution. Similarly, the evaluation method's difference, where the built-in AUC measure for the Keras package assesses AUC scores based on all prediction values rather than solely the "MEL" category, might inflate train and validation scores.
2) **Dataset quality and incidence rates:** Evaluation data scores fall below test scores, potentially due to qualitative differences in datasets. The Kaggle datasets are thoroughly

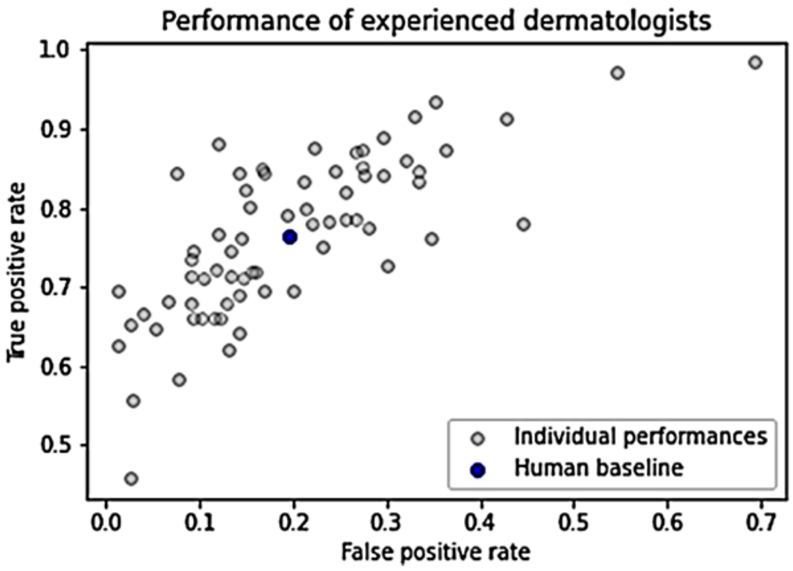

Figure 1 **Dermatologists' performance.** Full-size DOI: 10.7717/peerj-cs.2795/fig-1

Table 6 **Overview of the different models.**

ID	Labels	Depth	Dropout	Train_AUC	Val_AUC	Test_AUC	Eval_AUC
1.	Binary	Deep	0	**0.995**	0.940	0.868	0.806
2.	Binary	Deep	0.2	0.99	0.948	0.895	0.766
3.	Binary	Deep	0.4	0.987	0.931	0.807	0.801
4.	Binary	Deep	0.6	0.988	0.946	0.883	0.746
5.	Binary	Shallow	0	0.971	0.912	0.751	0.705
6.	Binary	Shallow	0.2	0.990	0.940	0.824	0.799
7.	Binary	Shallow	0.4	0.994	0.950	0.885	0.763
8.	Binary	Deep	0.6	0.990	0.950	0.886	0.795
9.	Multiclass	Deep	0	0.990	0.940	0.887	0.809
10.	Multiclass	Deep	0.2	0.986	**0.983**	0.549	0.593
11.	Multiclass	Deep	0.4	0.984	0.690	**0.897**	**0.822**
12.	Multiclass	Deep	0.6	0.981	0.981	0.885	0.762
13.	Multiclass	Shallow	0	0.981	0.927	0.766	0.762
14.	Multiclass	Shallow	0.2	0.978	0.934	0.713	0.747
15.	Multiclass	Shallow	0.4	0.986	0.981	0.863	0.737
16.	Multiclass	Shallow	0.6	0.983	0.980	0.886	0.737

Note:
The bold text indicates the highest score per category.

processed, potentially possessing higher quality than the evaluation images (*Banachewicz & Massaron, 2022*; *Khan et al., 2022*; *Esteva et al., 2019*). Additionally, the test dataset might bear a closer resemblance to the training images compared to the evaluation set, causing bias favoring test set performance. Furthermore, differences in incidence rates could contribute to this discrepancy.

3) Model 10 displays anomalous behavior. While training performance is robust, the remaining scores notably plummet. Examination of probability outcomes reveals similarities to the initial model. Visual inspection of AUC progression and loss during training indicates normal initial training, followed by a severe decline in performance—an occasional occurrence possibly stemming from stochastic weight initialization (*Sutskever et al., 2013*). The primary criterion for model selection centers on performance with novel data, specifically test and evaluation set performances. Under these criteria, Model 11 emerges as the superior performer. Notably, it demonstrates optimal performance on novel data while mirroring performance consistency across training and validation sets, indicating an absence of overfitting.

Visual inspection of TensorBoard output underscores a model rapidly learning training data but stabilizing after approximately 17 epochs, exhibiting consistent performance on both the AUC score and loss function between training and validation data. Thus, this model is deemed the baseline CNN model for this study.

On the test set, the baseline CNN attains:

- **AUC score:** AUC score of 0.897, placing it at the 37th percentile from the bottom, considerably distant from the winning score of 0.949 in the Kaggle competition field.
- **Evaluation dataset score:** AUC score of 0.822 on the evaluation dataset, surpassing the 0.773 scored by Tensorflow model.

Nonetheless, compared to human baseline performances, the CNN baseline falls short.

- **TPR comparison:** At comparable FPR rates, the average human exhibits TPR = 0.765, whereas the CNN records TPR = 0.694.
- **FPR comparison:** At analogous TPR rates, the average human demonstrates FPR = 0.196, while the CNN displays FPR = 0.320.

A visual comparison of these CNN models and the human average is depicted in Fig. 2. A juxtaposition between the final model and the initial training outcomes demonstrates a substantial enhancement in model performance and behavior. Notably, the final model comprises only 201,188 parameters, significantly fewer than the 11,840,195 parameters in the initial training model, as evident in Table 7. Additionally, Table 8 shows a broader range of prediction values across almost all categories in the final model. This discrepancy signifies the final model's capability to discern image features rather than merely learning the training distribution.

In summary, the baseline CNN converges effectively, outperforming the model (*Pang, Nijkamp & Wu, 2020*) but falling short of the average human performance in the study. The subsequent section will introduce and evaluate hybrid algorithms founded on predictions from both the human and CNN baselines.

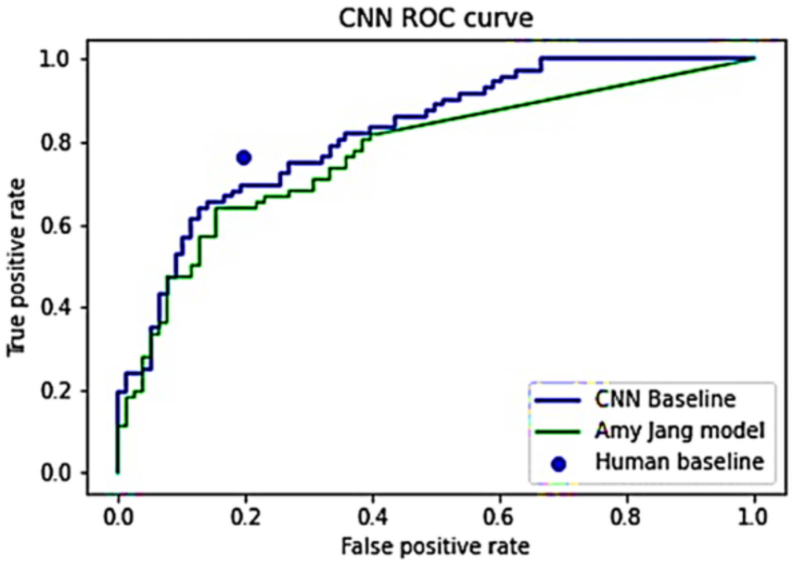

Figure 2 ROC curves for *Jang*'s *(2025)* model and CNN baseline model.
Full-size DOI: 10.7717/peerj-cs.2795/fig-2

Table 7 Model summary of the top-performing model.

Layer	Output shape	Number of parameters
EfficientNetb3 (Functional)	(None, 1,536)	10,783,535
dropout (Dropout)	(None, 1,536)	0
dense 6 (Dense)	(None, 128)	196,763
dense 7 (Dense)	(None, 128)	4,128
dense 8 (Dense)	(None, 9)	297

Table 8 Descriptive statistics of the probability estimates from the top-performing model on the test set.

Label	0	1	2	3	4	5	6	7	8
Min	4.5e−11	9.3e−12	3.8e−09	2.6e−11	3.9e−07	2.0e−6	1.4e−11	7.5e−08	1.1e−11
Median	5.5e−06	9.3e−12	6.1e−05	2.3e−06	6.6e−04	3.9e−3	1.8e−06	0.99	1.8e−06
Max	0.94	0.59	0.15	0.75	0.996	1.0	0.29	1.0	0.15

HYBRID ALGORITHMS

Augmented hybrid approach

In certain instances, image classification poses varying levels of difficulty, presenting disparities between human and machine capacities (*Kaluarachchi, Reis & Nanayakkara, 2021*). The study by *Han et al. (2018)* demonstrated discrepancies in image interpretation, where image challenges for human subjects exhibited notably high performance when processed by algorithms. Similarly, *Marchetti et al. (2020)* revealed enhancements in performance by replacing uncertain human responses with predictions generated by a

CNN. Defining a mutual relationship in the difficulty of image interpretation—where some images present challenges for humans but not for machines, while others are easily interpreted by humans yet pose complexity for computational systems (*Archana & Jeevaraj, 2024*). Moreover, it proposes leveraging CNN prediction values as indicators to delineate images deemed "easy" or "difficult" by the system (*Sun et al., 2023*). If substantiated, this insight could facilitate the construction of a synthesized list comprising both human and computer predictions, derived from the certainties inherent in CNN predictions (*Jackson et al., 2025*).

Algorithm

In the scenario presented in Eqs. (2)–(7), there exists an array denoted as A, comprising prediction values a_1 through a_n, where 'n' represents the count of assessed images. The task involves arranging this array in ascending order, leading to the formation of array B, consisting of elements b_1 through b_n. Simultaneously, it is imperative to maintain a clear correspondence or mapping between the index values of the original array A and the resulting array B. Introducing a parameter labeled as 's', let us define 'i' and 'j' as follows:

$$i = \left|\frac{n}{s}\right| \tag{2}$$

$$j = \left|\frac{(s-1)n}{s}\right|. \tag{3}$$

The constraint stipulates that 's' must be greater than 2 to uphold the condition where 'i' is less than 'j'. Subsequently, the partitioning of 'B' into three lists can be executed as:

$$B_{lower} = [b_1, b_2, \ldots, b_i] \tag{4}$$

$$B_{inner} = \left[b_{i+1}, b_{i+2}, \ldots, b_j\right] \tag{5}$$

$$B_{upper} = \left[b_{j+1}, b_{j+2}, \ldots, b_n\right]. \tag{6}$$

The merging of B_{lower} and B_{upper} is designated as B_{outer}.

$$B_{outer} = \left[b_1, b_2, \ldots, b_i, b_{j+1}, b_{j+2}, \ldots, b_n\right]. \tag{7}$$

The array denotes values extracted from B that demonstrate heightened "certainty" by their proximity to either 0 or 1. Conversely, the B_{inner} signifies values from B characterized by reduced certainty, specifically those closer to 0.5. Upon the establishment of both the internal and external lists, the indices corresponding to the images within one of these lists can be utilized as inputs for Algorithm 1, designated as the "Subset index list". Subsequently, this algorithm substitutes the CNN predictions linked to the Subset index list with human responses randomly chosen for those specific images (lines 4–11), thereby generating a substituted predictions list.

Following this, an analysis employing standard ROC/AUC methodology, as introduced in "Empirical Study", is conducted on this list. Due to the algorithm's random selection of human predictions for each index, stochasticity becomes inherent, resulting in varying

Algorithm 1 **Augmented hybrid approach.**

```
START
input: CNN predictions, Human predictions, Ground truths, Subset index list, Number of iterations,
  threshold list
output: ROC Curve, AUC score
1 ROC_Curves = 3DTensor
2 AUC_Scores = list
3 for each iteration do
4     Create substituted prediction list:
5     sub_predictions = list
6     for each index in CNN prediction do
7         if index in Subset index list then
8             Pick random human prediction for corresponding image
9             sub_predictions[index] ← humanpredictions[index]
10        else
11            sub_predictions[index] ← CNNpredictions[index]
12    for loop end
13 Create ROC Curves & AUC Scores:
14 ROC, AUC ← ROC/AUCAnalysis(input = (sub_predictions, Groundtruths, thresholdlist)
15 UpdateROC_Curves ← ROC (stacking the dataframes on top of each other)
16 UpdateAUC_scores ← AUC
17 for loop end
18 Generate average ROC Curve:
19 for each cell $c_{ijn}$ in ROC_curve do
20 |  $c_{ij} \leftarrow \frac{1}{n}\sum_{n=1}^{n} c_{ijn}$
21 for loop end
22 Plot = Lineplot showing Sensitivity and Specificity measures
23 Average AUC $\leftarrow \frac{1}{n}\sum_{n=1}^{n}$ AUC_scores
24 Return Plot, Average AUC
END
```

outcomes across different trials. Addressing this variability involves executing the setup multiple times, as per the tenet of the strong law of large numbers, which asserts that the anticipated result from an infinite number of trials will converge toward the population parameter. In the algorithm's specified lines (18–23), a sequence of 1,000 trials was executed, collecting ROC curves and AUC scores for each trial. These values were subsequently consolidated using averaging. The method employed for averaging involved

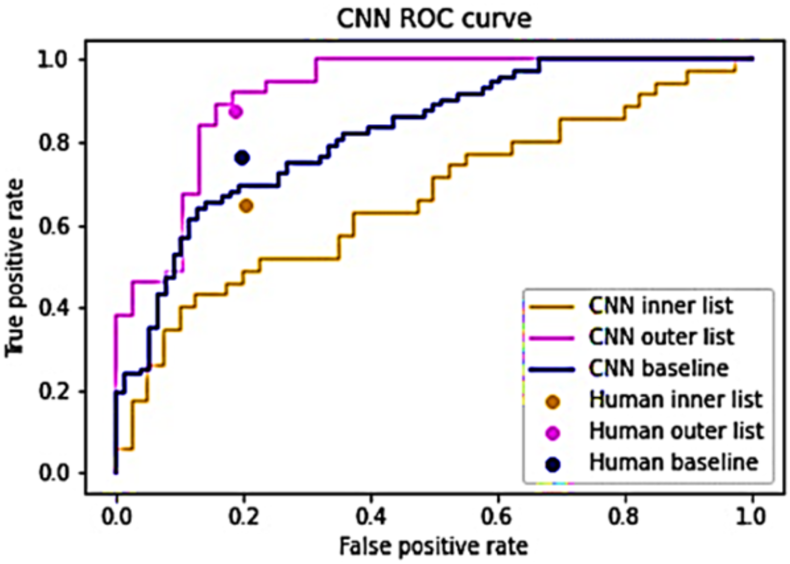

Figure 3 Performance of humans and CNN for the inner and outer sub-lists.
Full-size DOI: 10.7717/peerj-cs.2795/fig-3

treating individual ROC curves as distinct DataFrames, each composed of (FPR and TPR) pairs. These DataFrames were concatenated vertically, forming a 3D Tensor structure. In this arrangement, each cell could be denoted as c_{ijn}, where 'i' signifies the row index, 'j' represents the column index, and 'n' indicates the depth corresponding to the specific iteration number. The subsequent step involved computing the average values across the depth dimension, resulting in the derivation of a final DataFrame.

Testing differences between humans and CNNs

The hybrid algorithm operates under the premise that humans and CNNs exhibit varying proficiency in analyzing distinct images, with their performances showing minimal correlation (*Bozkurt, 2022*). A strong correlation exists between their performances, substituting one entity's prediction with the other should yield negligible or no impact. To validate this presumption, the CNN predictions were segregated into inner and outer lists following Eqs. (5) and (7). A value of $s = 4$ was employed to maintain equal list sizes (*Wang, Wong & Lu, 2020*; *Fawcett, 2006*). Both lists underwent evaluation by both the baseline CNN model and human evaluators. The outcomes, depicted in Fig. 3, highlight differential ease in identifying certain images. The CNN's performance appears least optimal for the inner list, registering an AUC of 0.664, while demonstrating superior performance on the outer list, yielding an AUC of 0.917. Similarly, human evaluators displayed lower performance on the inner list and higher performance on the outer list. Notably, their TPR exhibited variation between the lists, whereas the FPR remained relatively consistent. This substantiates the assertion that certain images pose greater classification challenges while suggesting the feasibility of utilizing CNN predictions to specifically target these challenging images (*Müller et al., 2024*). For the inner list, CNN's performance significantly trails behind human performance. Conversely, on the outer list, CNN's performance marginally surpasses human performance (*Mahmood et al., 2024*).

These findings corroborate the notion that humans and CNNs encounter difficulties in analyzing dissimilar images, thereby supporting the argument against a strong correlation between their performances (*Mahmood et al., 2024*).

Augmented hybrid results

After demonstrating variations in image difficulty for classification and the lack of a strong correlation between human and CNN performance, the subsequent phase involves ensembling human and CNN responses into a unified list (*Ganaie et al., 2022*; *Wu et al., 2022*). It is postulated that an arrangement where human answers constitute the inner elements and CNN predictions form the outer elements will outperform the baseline performances (*Akram et al., 2025*; *Archana & Jeevaraj, 2024*). To evaluate this hypothesis, Algorithm 1 is executed twice: once with the inner list as the "Subset index list" and once with the outer list in the same role. Both instances of the algorithm are run 1,000 times, with an expectation of convergence towards the anticipated ROC curve and AUC values. Figure 4 presents the outcomes derived from generating and scrutinizing the inner and outer substitution lists.

The outer substitution list broadly mirrors the CNN baseline model, deviating slightly at the extremities, showcasing an AUC score of 0.798—marginally lower than the baseline CNN score of 0.822 (*Archana & Jeevaraj, 2024*). In contrast, the inner substitution line exhibits a distinct pattern: below the CNN baseline within FPR [0, 0.17], surpassing the baseline within FPR [0.17, 0.4], and descending below the baseline within FPR [0.4, 1]. Its AUC score of 0.772 falls slightly beneath the CNN baseline. Nevertheless, the inner substitution line boasts a higher TPR (0.782) compared to the human baseline at similar FPR levels, while also demonstrating a lower FPR (0.182) compared to the human baseline at similar TPR values (*Shahid et al., 2025*; *Caruana & Niculescu-Mizil, 2006*).

STUDY LIMITATIONS

There are certain limitations of this study. Addressing these limitations could strengthen the robustness and applicability of the study's findings in real-world clinical practice.

Ethical considerations

The ethical considerations around having machine-driven decisions in life-critical medical situations using the pure augmented hybrid algorithm are not fully explored. Granting final decision authority to humans is mentioned as a solution but not fleshed out. This study is the first step toward a hybrid algorithm. Nonetheless, this could be addressed by granting a human final decision-making authority which is aligned with the global concept of human-in-the-loop (*Sasseville et al., 2025*; *Schuitmaker et al., 2025*; *Siddique et al., 2024*; *van den Berg, 2024*; *Topol, 2019*).

Limited dataset size and generalizability

The study's findings may not be directly applicable to all dermatological settings due to the specific dataset used, which consists of 150 images focused on particular characteristics of skin lesions (*Reddy et al., 2023*; *Shahid et al., 2025*; *Tschandl et al., 2019*). The small size of the evaluation dataset and the potential selection bias limits the generalizability of the

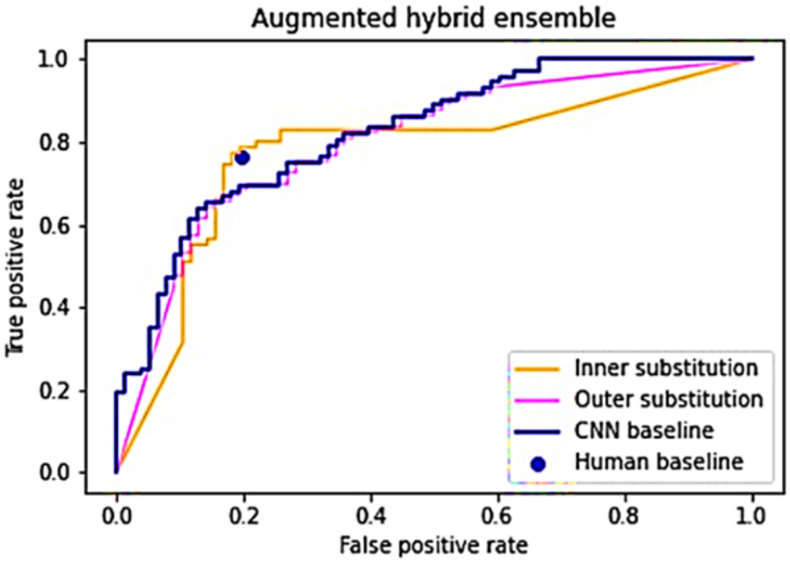

Figure 4 **An average of 1,000 simulations of the inner and outer hybrid lists.**
Full-size DOI: 10.7717/peerj-cs.2795/fig-4

results (*Sabazade et al., 2025*; *Esteva et al., 2017*). The study does not explicitly address whether the dataset fully represents real-world clinical scenarios (*Liu et al., 2024*; *Bejnordi et al., 2017*). Different populations, imaging techniques, or lesion types may yield different results (*Grzybowski, Jin & Wu, 2024*; *Winkler et al., 2019*; *Bejnordi et al., 2017*). A larger, more diverse dataset in future studies would help address these concerns and strengthen the conclusions (*Yan et al., 2025*; *Chen et al., 2024*; *Tschandl et al., 2020*).

Model complexity and interpretability

While the CNN models used, particularly EfficientNetB3, demonstrate strong performance, they inherently function as "black boxes," limiting interpretability (*Räz, 2024*). This lack of transparency is a critical concern in medical settings, where clinicians require not only accurate predictions but also insights into the reasoning behind them to build trust and validate results. Without interpretability, it becomes difficult to detect biases, troubleshoot errors, or confidently apply the model's predictions in high-stakes scenarios. Incorporating techniques such as saliency maps, attention mechanisms, or SHAP values could enhance transparency by identifying which features or regions of an image influence the model's output (*Cohen-Inger et al., 2025*; *Deng et al., 2024*; *Soomro, Niaz & Choi, 2024*). Hybrid approaches, combining interpretable rule-based models with deep learning, may also strike a balance between performance and explainability (*Khalil et al., 2023*; *Caruana & Niculescu-Mizil, 2006*). While this study focuses primarily on performance, interpretability remains a crucial area for future research to ensure reliable clinical integration (*Coots et al., 2025*; *Baumann et al., 2024*). Addressing the trade-off between model complexity and interpretability will be key to gaining practitioner trust and achieving better patient outcomes (*Guyton, Pak & Rovira, 2025*; *Bria, Marrocco & Tortorella, 2020*).

Inclusion of metadata

The decision to exclude metadata from the model may overlook potentially valuable information that could improve diagnostic accuracy (*Khan et al., 2025*; *Duan et al., 2024*; *Wang, Wong & Lu, 2020*). Incorporating metadata, such as patient age and lesion location, might enhance the model's performance (*Guermazi et al., 2024*; *Kania, Montecinos & Goldberg, 2024*; *Esteva et al., 2019*).

Resource intensive training

Training deep learning models, especially on large datasets with complex architectures, requires significant computational resources and time (*Rahman et al., 2021*; *Bria, Marrocco & Tortorella, 2020*; *Litjens et al., 2017*). This limitation could hinder the broader adoption and replication of the study's findings, particularly in resource-constrained settings (*Zhang et al., 2024a*, *2024b*; *Topol, 2019*).

Dependency on histopathological examination

The ground truth labels for the dataset were established *via* histopathological examination, which itself has limitations (*McCaffrey et al., 2024*; *Göndöcs & Dörfler, 2024*), including sampling bias (*Wang et al., 2024*; *Webb et al., 2024*) and interobserver variability (*Pinello et al., 2025*; *Shinde et al., 2025*). Reliance solely on histopathology may introduce errors in the dataset labels (*Zhang et al., 2024b*).

Human baseline variability

The human baseline performance was established using a subset of participants experienced in dermoscopy. However, individual variability in human performance (*Naseri & Safaei, 2025*; *Stevens et al., 2025*; *Naeem et al., 2024*), even among experienced dermatologists (*Gupta et al., 2025*; *Rubegni et al., 2024*), could introduce uncertainty (*Sanz-Motilva et al., 2024*) in comparison with the CNN models (*Miller et al., 2024*; *Ali et al., 2023*).

Dataset imbalance

The dataset used for training and evaluation may suffer from class imbalance issues (*Liu et al., 2020*; *He & Garcia, 2009*), particularly with the low incidence rate of malignant cases (*Gurcan & Soylu, 2024*). This imbalance could affect the model's performance and generalizability (*Fang et al., 2025*).

Exploring model enhancements for improved performance

This study demonstrates how a hybrid approach—combining CNN predictions with human expertise—outperforms individual baselines (*Nugroho, Ardiyanto & Nugroho, 2023*), including both standalone CNNs (*Liu et al., 2025*) and human assessments (*Selvaraj et al., 2024*; *Esteva et al., 2017*). However, the AUC scores achieved are lower than those reported in recent studies, such as *Houssein et al. (2024)* and *Nugroho, Ardiyanto & Nugroho (2023)*, which utilize more advanced CNN architectures and training techniques.

Future work will focus on testing our hybrid approach with these newer methods and exploring ways to adapt or incorporate them into our framework to achieve further performance gains. Benchmarking against these recent advancements will help ensure our approach remains competitive and aligned with the latest developments in skin lesion classification.

CONCLUSION

This research introduces a novel augmented hybrid approach that combines the strengths of CNNs with selective human intervention, aimed at enhancing skin lesion classification accuracy. By leveraging the EfficientNetB3 backbone, known for its balance between performance and efficiency, this study advances the field of medical image analysis with a focus on practicality and scalability (*Esteva et al., 2019*). The hybrid algorithm prioritizes high-confidence CNN predictions while delegating uncertain cases to medical experts, thereby optimizing diagnostic outcomes with minimal human resource expenditure.

Our comprehensive evaluation of the ISIC-2019 and ISIC-2020 datasets compared against 69 trained medical professionals demonstrates the promise of this approach (*ISIC, 2024*). The baseline CNN model achieved a competitive AUC score of 0.822, performing close to human experts. However, the hybrid model improved upon these results, achieving a TPR of 0.782 and reducing the FPR to 0.182, showcasing the effectiveness of combining human and machine intelligence. These findings underscore the practical potential of integrating CNNs into clinical workflows while ensuring that human expertise remains central to decision-making (*Rawat, Rajendran & Sikarwar, 2025*; *Gholizadeh, Rokni & Babaei, 2024*).

While the hybrid approach offers improved diagnostic accuracy and resource efficiency, challenges persist. Issues such as dataset imbalance, model interpretability, and computational resource demands highlight the need for further research to refine and generalize the methodology (*Strika et al., 2025*; *Char, Shah & Magnus, 2018*). The exclusion of metadata, though intentional in this study, also points to opportunities for future work that may enhance diagnostic performance by incorporating contextual clinical information (*Hermosilla et al., 2024*; *Jones et al., 2022*). Moreover, ethical considerations surrounding human-in-the-loop frameworks require careful attention to ensure that the technology serves as a support system, not a replacement, for clinical judgment (*Lee et al., 2025*).

This research contributes to the growing body of literature on AI-assisted diagnostics by demonstrating the potential of hybrid intelligence models to bridge the gap between human expertise and algorithmic efficiency. The results indicate that well-structured collaboration between CNNs and medical professionals can mitigate the limitations of both systems. Moving forward, this hybrid framework offers a scalable, pragmatic solution for clinical settings, fostering more reliable and accurate skin lesion diagnosis while efficiently managing healthcare resources.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

Deep Himmatbhai Ajabani is employed by Source InfoTech Inc. and Karar Ali is employed by VentureDive Pvt. Limited.

Author Contributions

- Deep Ajabani conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Zaffar Ahmed Shaikh conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Amr Yousef conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Karar Ali performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Marwan A. Albahar analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The ISIC-2019 and ISIC-2020 datasets are available at: https://www.isic-archive.com/.

The winning solution (*i.e.*, code and algorithm) of the 2019 SIIM-ISIC melanoma classification challenge (*Ha, Liu & Liu, 2020*) is available at GitHub (*GitHub, 2024a*) at https://github.com/haqishen/SIIM-ISIC-Melanoma-Classification-1st-Place-Solution.

The original code/algorithm of the 2020 SIIM-ISIC melanoma classification challenge winner is available on GitHub: https://github.com/ISIC-Research/ADAE.

The BCN20000 dataset of the 19,424 images of skin lesions captured from 2010 to 2016 of the Three-Point Checklist of Dermoscopy (*Combalia et al., 2019*) is available at arXiv: https://doi.org/10.48550/arXiv.1908.02288.

The training data for the ISIC-2019 dataset is available at Kaggle: https://www.kaggle.com/datasets/andrewmvd/isic-2019.

The original melanoma skin cancer dataset of 10,000 images is available at Kaggle: https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images.

REFERENCES

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker PA, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. 2016. TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016)*. Savannah, GA, USA, 265–283.

Abbas S, Ahmed F, Khan WA, Ahmad M, Khan MA, Ghazal TM. 2025. Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence. *Scientific Reports* **15(1)**:1746 DOI 10.1038/s41598-024-83966-4.

Abdelrahman A, Viriri S. 2023. EfficientNet family U-Net models for deep learning semantic segmentation of kidney tumors on CT images. *Frontiers in Computer Science* **5**:1235622 DOI 10.3389/fcomp.2023.1235622.

Aboulmira A, Hrimech H, Lachgar M, Hanine M, Garcia CO, Mezquita GM, Ashraf I. 2025. Hybrid model with wavelet decomposition and EfficientNet for accurate skin cancer classification. *Journal of Cancer* **16(2)**:506–520 DOI 10.7150/jca.101574.

Adegun A, Viriri S. 2021. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review* **54(2)**:811–841 DOI 10.1007/s10462-020-09865-y.

Ahmad H, Shahab S, Mobarak WF, Dutta AK, Abolelmagd YM, Shaikh ZA, Anjum M. 2024. Convergence results for cyclic-orbital contraction in a more generalized setting with application. *AIMS Mathematics* **9(6)**:15543–15558 DOI 10.3934/math.2024751.

Akhund TMNU, Ajabani D, Shaikh ZA, Elrashidi A, Nureldeen WA, Bhatti MI, Sarker MM. 2024a. A comprehensive exploration of human communal media interaction and its evolving impact on psychological health across demographics and time. *PeerJ Computer Science* **10(6)**:e2398 DOI 10.7717/peerj-cs.2398.

Akhund TMNU, Shaikh ZA, De La Torre Díez I, Gafar M, Ajabani DH, Alfarraj O, Tolba A, Fabian-Gongora H, López LAD. 2024b. IoST-enabled robotic arm control and abnormality prediction using minimal flex sensors and Gaussian mixture models. *IEEE Access* **12(6)**:45265–45278 DOI 10.1109/ACCESS.2024.3380360.

Akram M, Adnan M, Ali SF, Ahmad J, Yousef A, Alshalali TAN, Shaikh ZA. 2025. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches. *Scientific Reports* **15(1)**:1342 DOI 10.1038/s41598-024-84478-x.

Alam TM, Shaukat K, Khan WA, Hameed IA, Almuqren LA, Raza MA, Aslam M, Luo S. 2022. An efficient deep learning-based skin cancer classifier for an imbalanced dataset. *Diagnostics* **12(9)**:2115 DOI 10.3390/diagnostics12092115.

Alhichri H, Alswayed AS, Bazi Y, Ammour N, Alajlan NA. 2021. Classification of remote sensing images using EfficientNet-B3 CNN model with attention. *IEEE Access* **9**:14078–14094 DOI 10.1109/ACCESS.2021.3051085.

Ali MU, Khalid M, Alshanbari H, Shaikh ZA, Lee SW. 2023. Enhancing skin lesion detection: a multistage multiclass convolutional neural network-based framework. *Bioengineering* **10(12)**:1430 DOI 10.3390/bioengineering10121430.

Ali MS, Miah MS, Haque J, Rahman MM, Islam MK. 2021. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications* **5**:100036 DOI 10.1016/j.mlwa.2021.100036.

Ali K, Shaikh ZA, Khan AA, Laghari AA. 2022. Multiclass skin cancer classification using EfficientNets—a first step towards preventing skin cancer. *Neuroscience Informatics* **2(4)**:100034 DOI 10.1016/j.neuri.2021.100034.

Alotaibi A, AlSaeed D. 2025. Skin cancer detection using transfer learning and deep attention mechanisms. *Diagnostics* **15(1)**:99 DOI 10.3390/diagnostics15010099.

Archana R, Jeevaraj PE. 2024. Deep learning models for digital image processing: a review. *Artificial Intelligence Review* **57(1)**:11 DOI 10.1007/s10462-023-10631-z.

**Argenziano G, Soyer HP, Chimenti S, Talamini R, Corona R, Sera F, Binder M, Cerroni L, Rosa GD, Ferrara G, Hofmann-Wellenhof R, Landthaler M, Menzies SW, Pehamberger H, Piccolo

D, Rabinovitz HS, Schiffner R, Staibano S, Stolz W, Bartenjev I, Blum A, Braun R, Cabo H, Carli P, Giorgi VD, Fleming MG, Grichnik JM, Grin CA, Halpern AC, Johr R, Katz B, Kenet RO, Kittler H, Kreusch J, Malvehy J, Mazzocchetti G, Oliviero M, Özdemir F, Peris K, Perotti R, Perusquia A, Pizzichetta MA, Puig S, Rao B, Rubegni P, Saida T, Scalvenzi M, Seidenari S, Stanganelli I, Tanaka M, Westerhoff K, Wolf IH, Braun-Falco O, Kerl H, Nishikawa T, Wolff K, Kopf AW. 2003. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. *Journal of the American Academy of Dermatology* **48(5)**:679–693 DOI 10.1067/mjd.2003.281.

Banachewicz K, Massaron L. 2022. *The Kaggle Book: data analysis and machine learning for competitive data science*. Birmingham: Packt Publishing Ltd.

Batool A, Byun YC. 2023. Lightweight EfficientNetB3 model based on depthwise separable convolutions for enhancing classification of leukemia white blood cell images. *IEEE Access* **11**:37203–37215 DOI 10.1109/ACCESS.2023.3266511.

Baumann J, Sapiezynski P, Heitz C, Hannák A. 2024. Fairness in online ad delivery. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. New York: ACM, 1418–1432.

Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, CAMELYON16 Consortium. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *The Journal of the American Medical Association* **318(22)**:2199–2210 DOI 10.1001/jama.2017.14585.

Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* **13(1)**:281–305 DOI 10.5555/2188385.2188395.

Bingol H, Alatas B. 2021. Classification of brain tumor images using deep learning methods. *Turkish Journal of Science and Technology* **16(1)**:137–143.

Bozkurt F. 2022. A comparative study on classifying human activities using classical machine and deep learning methods. *Arabian Journal for Science and Engineering* **47(2)**:1507–1521 DOI 10.1007/s13369-021-06008-5.

Bozkurt F. 2023. Skin lesion classification on dermatoscopic images using effective data augmentation and pre-trained deep learning approach. *Multimedia Tools and Applications* **82(12)**:18985–19003 DOI 10.1007/s11042-022-14095-1.

Brady A, Neri E. 2020. Artificial intelligence in radiology—ethical considerations and challenges. *The British Journal of Radiology* **93(1108)**:20190667 DOI 10.3390/diagnostics10040231.

Bria A, Marrocco C, Tortorella F. 2020. Addressing class imbalance in deep learning for small lesion detection on medical images. *Computers in Biology and Medicine* **120(132)**:103735 DOI 10.1016/j.compbiomed.2020.103735.

Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, Weichenthal M, Klode J, Schadendorf D, Holland-Letz T, von Kalle C, Fröhling S, Schilling B, Utikal JS. 2019. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer* **119(10151)**:11–17 DOI 10.1016/j.ejca.2019.05.023.

Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, Berking C, Steeb T, Enk AH, von Kalle C, Von Kalle C. 2018. Skin cancer classification using convolutional neural networks: systematic review. *Journal of Medical Internet Research* **20(10)**:e11936 DOI 10.2196/11936.

Caruana R, Niculescu-Mizil A. 2006. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York: ACM, 161–168.

Cassidy B, Kendrick C, Brodzicki A, Jaworek-Korjakowska J, Yap MH. 2022. Analysis of the ISIC image datasets: usage, benchmarks and recommendations. *Medical Image Analysis* **75(4)**:102305 DOI 10.1016/j.media.2021.102305.

Char DS, Shah NH, Magnus D. 2018. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine* **378(11)**:981–983 DOI 10.1056/NEJMp1714229.

Chatterjee S, Gil JM, Byun YC. 2024. Early detection of multiclass skin lesions using transfer learning-based IncepX-ensemble model. *IEEE Access* **12**:113677–113693 DOI 10.1109/ACCESS.2024.3432904.

Chen Y, Wen Z, Chen J, Huang J. 2024. Enhancing the performance of bandit-based hyperparameter optimization. In: *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. Piscataway: IEEE, 967–980.

Chollet F, Chollet F. 2021. *Deep learning with Python. Simon and Schuster*. Second Edition. Shelter Island, NY, United States: Manning Publications Co. LLC, 2021. *Available at https://www.manning.com/books/deep-learning-with-python-second-edition* (accessed 27 August 2024).

Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H, Halpern A. 2018. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Piscataway: IEEE, 168–172.

Cohen-Inger N, Cohen S, Rabaev N, Rokach L, Shapira B. 2025. BiasGuard: guardrailing fairness in machine learning production systems. ArXiv preprint DOI 10.48550/arXiv.2501.04142.

Combalia M, Codella NC, Rotemberg V, Helba B, Vilaplana V, Reiter O, Carrera C, Barreiro A, Halpern AC, Puig S, Malvehy J. 2019. BCN20000: dermoscopic lesions in the wild. ArXiv preprint DOI 10.48550/arXiv.1908.02288.

Coots M, Linn KA, Goel S, Navathe AS, Parikh RB. 2025. Racial bias in clinical and population health algorithms: a critical review of current debates. *Annual Review of Public Health* **46**:112058 DOI 10.1146/annurev-publhealth-071823-112058.

Davis J, Goadrich M. 2006. The relationship between precision-recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. Pittsburgh, Pennsylvania, 233–240.

Dayananda C, Yamanakkanavar N, Nguyen T, Lee B. 2023. AMCC-Net: an asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images. *Engineering Applications of Artificial Intelligence* **122(22)**:106154 DOI 10.1016/j.engappai.2023.106154.

De A, Mishra N, Chang HT. 2024. An approach to the dermatological classification of histopathological skin images using a hybridized CNN-DenseNet model. *PeerJ Computer Science* **10(Pt 5)**:e1884 DOI 10.7717/peerj-cs.1884.

Debelee TG. 2023. Skin lesion classification and detection using machine learning techniques: a systematic review. *Diagnostics* **13(19)**:3147 DOI 10.3390/diagnostics13193147.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 248–255.

Deng S, Zhu Y, Yu Y, Huang X. 2024. An integrated approach of ensemble learning methods for stock index prediction using investor sentiments. *Expert Systems with Applications* **238(2019)**:121710 DOI 10.1016/j.eswa.2023.121710.

Deotte C. 2020. Triple stratified K fold with TF Records. In: *Kaggle: SIIM-ISIC-Melanoma Classification*. California, United States: Kaggle. *Available at https://www.kaggle.com/competitions/siim-isic-melanoma-classification/discussion/169139* (Assessed 27 August 2024).

Dietterich T. 1995. Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)* **27(3)**:326–327 DOI 10.1145/212094.212114.

Dillon JV, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, Patton B, Alemi A, Hoffman M, Saurous RA. 2017. TensorFlow distributions. ArXiv preprint DOI 10.48550/arXiv.1711.10604.

Duan J, Xiong J, Li Y, Ding W. 2024. Deep learning based multimodal biomedical data fusion: an overview and comparative review. *Information Fusion* **112(9)**:102536 DOI 10.1016/j.inffus.2024.102536.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542(7639)**:115–118 DOI 10.1038/nature21056.

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. 2019. A guide to deep learning in healthcare. *Nature Medicine* **25(1)**:24–29 DOI 10.1038/s41591-018-0316-z.

Fang X, Easwaran A, Genest B, Suganthan PN. 2025. Your data is not perfect: towards cross-domain out-of-distribution detection in class-imbalanced data. *Expert Systems with Applications* **267(3)**:126031 DOI 10.1016/j.eswa.2024.126031.

Farea E, Saleh RA, AbuAlkebash H, Farea AA, Al-antari MA. 2024. A hybrid deep learning skin cancer prediction framework. *Engineering Science and Technology, an International Journal* **57(4)**:101818 DOI 10.1016/j.jestch.2024.101818.

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* **27(8)**:861–874 DOI 10.1016/j.patrec.2005.10.010.

Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, Pirracchio R. 2022. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digital Medicine* **5(1)**:66 DOI 10.1038/s41746-022-00611-y.

Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, Jain A, Walter FM, Williams HC, Deeks JJ. 2020. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* **368**:m127 DOI 10.1136/bmj.m127.

Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. 2022. Ensemble deep learning: a review. *Engineering Applications of Artificial Intelligence* **115**:105151 DOI 10.1016/j.engappai.2022.105151.

Gholizadeh N, Rokni GR, Babaei M. 2024. Advantages and Disadvantages of using AI in dermatology. *Dermatological Reviews* **5(4)**:e248 DOI 10.1002/der2.248.

GitHub. 2024a. SIIM-ISIC melanoma classification 1st place solution. *Available at https://github.com/haqishen/SIIM-ISIC-Melanoma-Classification-1st-Place-Solution* (accessed 27 August 2024).

GitHub. 2024b. Winning model of the SIIM/ISIC 2020 grand challenge. *Available at https://github.com/ISIC-Research/ADAE* (accessed 27 August 2024).

Goceri E. 2020. Image augmentation for deep learning based lesion classification from skin images. In: *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*. Piscataway: IEEE, 144–148.

Gómez-Carmona O, Casado-Mansilla D, López-de-Ipiña D, García-Zubia J. 2024. Human-in-the-loop machine learning: reconceptualizing the role of the user in interactive approaches. *Internet of Things* **25**:101048 DOI 10.1016/j.iot.2023.101048.

Göndöcs D, Dörfler V. 2024. AI in medical diagnosis: AI prediction & human judgment. *Artificial Intelligence in Medicine* **149(3)**:102769 DOI 10.1016/j.artmed.2024.102769.

Goodfellow I. 2016. *Deep learning*. Cambridge, MA, United States: The MIT Press, 2–5. *Available at https://mitpress.mit.edu/9780262035613/deep-learning/* (accessed 27 August 2024).

Gouda W, Sama NU, Al-Waakid G, Humayun M, Jhanjhi NZ. 2022. Detection of skin cancer based on skin lesion images using deep learning. *Healthcare* **10(7)**:1183 DOI 10.3390/healthcare10071183.

Grzybowski A, Jin K, Wu H. 2024. Challenges of artificial intelligence in medicine and dermatology. *Clinics in Dermatology* **42(3)**:210–215 DOI 10.1016/j.clindermatol.2023.12.013.

Guermazi D, Shah A, Yumeen S, Vance T, Saliba E. 2024. Skinformatics: navigating the big data landscape of dermatology. *Journal of the European Academy of Dermatology and Venereology* **38(12)**:2217–2224 DOI 10.1111/jdv.20319.

Gulli A, Pal S. 2017. *Deep learning with Keras*. Birmingham: Packt Publishing Ltd.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316(22)**:2402–2410 DOI 10.1001/jama.2016.17216.

Gupta J, Sibbald C, Weinstein M, Pusic M, Bell M, MacLellan N, Bobotsis R, Brar R, Boutis K. 2025. Rash decisions: improving pediatrician skills in dermatologic diagnosis. *The Journal of Pediatrics* **278**:114436 DOI 10.1016/j.jpeds.2024.114436.

Gurcan F, Soylu A. 2024. Learning from imbalanced data: integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers* **16(19)**:3417 DOI 10.3390/cancers16193417.

Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A. 2016. Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). ArXiv preprint DOI 10.48550/arXiv.1605.01397.

Guyton Z, Pak R, Rovira E. 2025. The role of automation etiquette and task-criticality on performance, workload, automation reliance, and user confidence. *Applied Ergonomics* **125**:104430 DOI 10.1016/j.apergo.2024.104430.

Ha Q, Liu B, Liu F. 2020. Identifying melanoma images using EfficientNet ensemble: winning solution to the SIIM-ISIC melanoma classification challenge. ArXiv preprint DOI 10.48550/arXiv.2010.05351.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, 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 DOI 10.1093/annonc/mdy166.

Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, Hofmann-Wellenhof R, Lallas A, Emmert S, Buhl T, Zutt M, Blum A, Abassi MS, Thomas L, Tromme I, Tschandl P, Enk A, Rosenberger A, Alt C, Bachelerie M, Bajaj S, Balcere A, Baricault S, Barthaux C, Beckenbauer Y, Bertlich I, Blum A, Bouthenet M, Brassat S, Buck PM, Buder-Bakhaya K, Cappelletti M, Chabbert C, Labarthe JD, DeCoster E, Deinlein T, Dobler M, Dumon D, Emmert S, Gachon-Buffet J, Gusarov M, Hartmann F, Hartmann J, Herrmann A, Hoorens I, Hulstaert E, Karls R, Kolonte A, Kromer C, Lallas A, Vasseux CLB, Levy-Roy A, Majenka P, Marc M, Bourret VM, Michelet-Brunacci N, Mitteldorf C, Paroissien J, Picard C, Plise D, Reymann V, Ribeaudeau F, Richez P, Plaine HR, Salik D, Sattler E, Schäfer S, Schneiderbauer R, Secchi T, Talour K, Trennheuser L, Wald A, Wölbing P, Zukervar P. 2020. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Annals of Oncology* **31(1)**:137–143 DOI 10.1016/j.annonc.2019.10.013.

Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, Beltraminelli H, Berking C, Betz-Stablein B, Blum A, Braun SA, Carr R, Combalia M, Fernandez-Figueras M, Ferrara G, Fraitag S, French LE, Gellrich FF, Ghoreschi K, Goebeler M, Guitera P, Haenssle HA, Haferkamp S, Heinzerling L, Heppt MV, Hilke FJ, Hobelsberger S, Krahl D, Kutzner H, Lallas A, Liopyris K, Llamas-Velasco M, Malvehy J, Meier F, Müller CSL, Navarini AA, Navarrete-Dechent C, Perasole A, Poch G, Podlipnik S, Requena L, Rotemberg VM, Saggini A, Sangueza OP, Santonja C, Schadendorf D, Schilling B, Schlaak M, Schlager JG, Sergon M, Sondermann W, Soyer HP, Starz H, Stolz W, Vale E, Weyers W, Zink A, Krieghoff-Henning E, Kather JN, von Kalle C, Lipka DB, Fröhling S, Hauschild A, Kittler H, Brinker TJ. 2021. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer* **156(5)**:202–216 DOI 10.1016/j.ejca.2021.06.049.

Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. 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 DOI 10.1016/j.jid.2018.01.028.

Han S, Mao H, Dally WJ. 2015. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. ArXiv preprint DOI 10.48550/arXiv.1510.00149.

Han SS, Park I, Chang SE, Lim W, Kim MS, Park GH, Chung KY, Na JI. 2020. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* **140(9)**:1753–1761 DOI 10.1016/j.jid.2020.01.036.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143(1)**:29–36 DOI 10.1148/radiology.143.1.7063747.

Hasan MK, Elahi MTE, Alam MA, Jawad MT, Martí R. 2022. DermoExpert: skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation. *Informatics in Medicine Unlocked* **28**:100819 DOI 10.1016/j.imu.2021.100819.

Hasan MR, Fatemi MI, Monirujjaman Khan M, Kaur M, Zaguia A. 2021. Comparative analysis of skin cancer (benign vs. malignant) detection using convolutional neural networks. *Journal of Healthcare Engineering* **2021(1)**:5895156 DOI 10.1155/2021/5895156.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd Edition. New York: Springer. *Available at https://link.springer.com/book/10.1007/978-0-387-84858-7*.

He H, Garcia EA. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21(9)**:1263–1284 DOI 10.1109/TKDE.2008.239.

Hekler A, Kather JN, Krieghoff-Henning E, Utikal JS, Meier F, Gellrich FF, Brinker TJ. 2020. Effects of label noise on deep learning-based skin cancer classification. *Frontiers in Medicine* **7**:177 DOI 10.3389/fmed.2020.00177.

Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, Berking C, Haferkamp S, Klode J, Schadendorf D, Schilling B, Holland-Letz T, Izar B, von Kalle C, Fröhling S, Brinker TJ, Schmitt L, Peitsch WK, Hoffmann F, Becker JC, Drusio C, Jansen P, Klode J, Lodde G, Sammet S, Schadendorf D, Sondermann W, Ugurel S, Zader J, Enk A, Salzmann M, Schäfer S, Schäkel K, Winkler J, Wölbing P, Asper H, Bohne A, Brown V, Burba B, Deffaa S, Dietrich C, Dietrich M, Drerup KA, Egberts F, Erkens A, Greven S, Harde V, Jost M, Kaeding M, Kosova K, Lischner S, Maagk M, Messinger AL, Metzner M, Motamedi R, Rosenthal A, Seidl U, Stemmermann J, Torz K, Velez JG, Haiduk J, Alter M, Bär C,

Bergenthal P, Gerlach A, Holtorf C, Karoglan A, Kindermann S, Kraas L, Felcht M, Gaiser MR, Klemke C, Kurzen H, Leibing T, Müller V, Reinhard RR, Utikal J, Winter F, Berking C, Eicher L, Hartmann D, Heppt M, Kilian K, Krammer S, Lill D, Niesert A, Oppel E, Sattler E, Senner S, Wallmichrath J, Wolff H, Gesierich A, Giner T, Glutsch V, Kerstan A, Presser D, Schrüfer P, Schummer P, Stolze I, Weber J, Drexler K, Haferkamp S, Mickler M, Stauner CT, Thiem A. 2019. Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer* **120**:114–121 DOI 10.1016/j.ejca.2019.07.019.

Hermosilla P, Soto R, Vega E, Suazo C, Ponce J. 2024. Skin cancer detection and classification using neural network algorithms: a systematic review. *Diagnostics* **14(4)**:454 DOI 10.3390/diagnostics14040454.

Hernández-Pérez C, Combalia M, Podlipnik S, Codella NC, Rotemberg V, Halpern AC, Reiter O, Carrera C, Barreiro A, Helba B, Puig S, Vilaplana V, Malvehy J. 2024. Bcn20000: dermoscopic lesions in the wild. *Scientific Data* **11(1)**:641 DOI 10.1038/s41597-024-03387-w.

Hosseinzadeh M, Hussain D, Zeki Mahmood FM, Alenizi FA, Varzeghani AN, Asghari P, Darwesh A, Malik MH, Lee S-W. 2024. A model for skin cancer using combination of ensemble learning and deep learning. *PLOS ONE* **19(5)**:e0301275 DOI 10.1371/journal.pone.0301275.

Houssein EH, Abdelkareem DA, Hu G, Hameed MA, Ibrahim IA, Younan M. 2024. An effective multiclass skin cancer classification approach based on deep convolutional neural network. *Cluster Computing* **27(9)**:12799–12819 DOI 10.1007/s10586-024-04540-1.

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. Honolulu, HI, USA, 4700–4708 DOI 10.1109/CVPR.2017.243.

Huang C, Wang W, Zhang X, Wang SH, Zhang YD. 2022. Tuberculosis diagnosis using deep transferred EfficientNet. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20(5)**:2639–2646 DOI 10.1109/TCBB.2022.3199572.

Hussain F, Nauman M, Alghuried A, Alhudhaif A, Akhtar N. 2023. Leveraging big data analytics for enhanced clinical decision-making in healthcare. *IEEE Access* **11**:127817–127836 DOI 10.1109/ACCESS.2023.3332030.

ISIC. 2024. International skin imaging collaboration. *Available at https://www.isic-archive.com/* (accessed 2 January 2024).

Jackson J, Jackson LE, Ukwuoma CC, Kissi MD, Oluwasanmi A, Qin Z. 2025. A patch-based deep learning framework with 5-B network for breast cancer multi-classification using histopathological images. *Engineering Applications of Artificial Intelligence* **148**:110439 DOI 10.1016/j.engappai.2025.110439.

Jafar MN, Ullah F, Idris SA, Mobarak WFM, Shaikh ZA, Chavali D, Muniba K. 2024. Evaluation of tourist destinations carrying capacity in a decision-making context with muirhead means aggregation operator in q-rung orthopair fuzzy hypersoft environment. *IEEE Access* **12**:162476–162498 DOI 10.1109/ACCESS.2024.3482391.

Jang A. 2025. TensorFlow + Transfer learning: melanoma. *Available at https://www.kaggle.com/code/amyjang/tensorflow-transfer-learning-melanoma/notebook* (accessed 27 August 2024).

Jones OT, Matin RN, Van der Schaar M, Bhayankaram KP, Ranmuthu CKI, Islam MS, Behiyat D, Boscott R, Calanzani N, Emery J, Williams HC, Walter FM. 2022. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *The Lancet Digital Health* **4(6)**:e466–e476 DOI 10.1016/S2589-7500(22)00023-1.

Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, Yoon DH. 2017. In-datacenter performance analysis of a tensor processing unit. In: *Proceedings of the 44th Annual*

International Symposium on Computer Architecture (ISCA), 1–12 DOI 10.1145/3079856.3080246.

Kaluarachchi T, Reis A, Nanayakkara S. 2021. A review of recent deep learning approaches in human-centered machine learning. *Sensors* **21(7)**:2514 DOI 10.3390/s21072514.

Kang M, Tian J. 2018. Machine learning: data preprocessing. In: Pecht MG, Kang M, eds. *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*. Hoboken: John Wiley & Sons, 111–130 DOI 10.1002/9781119515326.ch5.

Kania B, Montecinos K, Goldberg DJ. 2024. Artificial intelligence in cosmetic dermatology. *Journal of Cosmetic Dermatology* **23(10)**:3305–3311 DOI 10.1111/jocd.16538.

Kassani SH, Kassani PH. 2019. A comparative study of deep learning architectures on melanoma detection. *Tissue and Cell* **58**:76–83 DOI 10.1016/j.tice.2019.04.009.

Kassem MA, Hosny KM, Damaševičius R, Eltoukhy MM. 2021. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. *Diagnostics* **11(8)**:1390 DOI 10.3390/diagnostics11081390.

Kay J, Schoels MM, Dörner T, Emery P, Kvien TK, Smolen JS, Breedveld FC. 2018. Consensus-based recommendations for the use of biosimilars to treat rheumatological diseases. *Annals of the Rheumatic Diseases* **77(2)**:165–174 DOI 10.1136/annrheumdis-2017-211937.

Keerthana D, Venugopal V, Nath MK. 2023. Hybrid convolutional neural networks with SVM classifier for classification of skin cancer. *Biomedical Engineering Advances* **3**:100045 DOI 10.1016/j.bea.2022.100069.

Ketkar N, Santana E. 2017. *Deep learning with Python.* Vol. 1. Berkeley: CA Press.

Khalil M, Naeem A, Naqvi RA, Zahra K, Muqarib SA, Lee SW. 2023. Deep learning-based classification of abrasion and ischemic diabetic foot sores using camera-captured images. *Mathematics* **11(17)**:3793 DOI 10.3390/math11173793.

Khamparia A, Singh PK, Rani P, Samanta D, Khanna A, Bhushan B. 2021. An Internet of Health things-driven deep learning framework for detection and classification of skin cancer using transfer learning. *Transactions on Emerging Telecommunications Technologies* **32(7)**:e3963 DOI 10.1002/ett.3963.

Khan SUR, Asif S, Zhao M, Zou W, Li Y, Li X. 2025. Optimized deep learning model for comprehensive medical image analysis across multiple modalities. *Neurocomputing* **619**:129182 DOI 10.1016/j.neucom.2024.129182.

Khan AA, Laghari AA, Shaikh AA, Shaikh ZA, Jumani AK. 2022. Innovation in multimedia using IoT systems. In: *Multimedia Computing Systems and Virtual Reality*. Boca Raton: CRC Press, 171–187.

Khan RH, Salamat N, Baig AQ, Shaikh ZA, Yousef A. 2024. Graph-based analysis of DNA sequence comparison in closed cotton species: a generalized method to unveil genetic connections. *PLOS ONE* **19(9)**:e0306608 DOI 10.1371/journal.pone.0306608.

Kim M, Bae HJ. 2020. Data augmentation techniques for deep learning-based medical image analyses. *Journal of the Korean Society of Radiology* **81(6)**:1290–1304 DOI 10.3348/jksr.2020.81.6.1290.

Kiziloluk S, Yildirim M, Bingol H, Alatas B. 2024. Multi-feature fusion and dandelion optimizer based model for automatically diagnosing the gastrointestinal diseases. *PeerJ Computer Science* **10**:e1919 DOI 10.7717/peerj-cs.1919.

Koçak B, Ponsiglione A, Stanzione A, Bluethgen C, Santinha J, Ugga L, Huisman M, Klontzas ME, Cannella R, Cuocolo R. 2024. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology* **31(2)**:75–88 DOI 10.4274/dir.2024.242854.

Kousar A, Ahmad J, Ijaz K, Yousef A, Shaikh ZA, Khosa I, Chavali D, Anjum M. 2024. MLHS-CGCapNet: a lightweight model for multilingual hate speech detection. *IEEE Access* **12**:106631–106644 DOI 10.1109/ACCESS.2024.3434664.

Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, Eriksson H, Lykou A, Linos E. 2024. Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis. *NPJ Digital Medicine* **7(1)**:78 DOI 10.1038/s41746-024-01031-w.

Kumar V, Prabha C, Sharma P, Mittal N, Askar SS, Abouhawwash M. 2024. Unified deep learning models for enhanced lung cancer prediction with ResNet-50-101 and EfficientNet-B3 using DICOM images. *BMC Medical Imaging* **24(1)**:63 DOI 10.1186/s12880-024-01241-4.

Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Carney PA, Bogart A, Argenziano G, Zalaudekg I, Wolf M. 2021a. Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences of the United States of America* **118(7)**:e2016884118 DOI 10.1073/pnas.1601827113.

Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Moussaid M, Argenziano G, Zalaudek I, Carney PA, Wolf M. 2019. How to detect high-performing individuals and groups: decision similarity predicts accuracy. *Science Advances* **5(11)**:eaaw9011 DOI 10.1126/sciadv.aaw9011.

Kurvers RHJM, Herzog SM, Hertwig R, Krause J, Wolf M. 2021b. Pooling decisions decreases variation in response bias and accuracy. *iScience* **24(7)**:102740 DOI 10.1016/j.isci.2021.102740.

Lane H, Sarkies M, Martin J, Haines T. 2017. Equity in healthcare resource allocation decision making: a systematic review. *Social Science & Medicine* **175**:11–27 DOI 10.1016/j.socscimed.2016.12.012.

LeCun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86(11)**:2278–2324 DOI 10.1109/5.726791.

Lee AKW, Chan LKW, Lee CH, Bohórquez JMC, Haykal D, Wan J, Yi KH. 2025. Artificial intelligence application in diagnosing, classifying, localizing, detecting and estimation the severity of skin condition in aesthetic medicine: a review. *Dermatological Reviews* **6**:e70015 DOI 10.1002/der2.70015.

Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. 2018. Hyperband: a novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research* **18(185)**:1–52.

Li W, Zhuang J, Wang R, Zhang J, Zheng WS. 2020. Fusing metadata and dermoscopy images for skin disease diagnosis. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Piscataway: IEEE, 1996–2000 DOI 10.1109/ISBI45749.2020.9098516.

Liaw LCM, Tan SC, Goh PY, Lim CP. 2025. A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification. *Information Sciences* **686**:121193 DOI 10.1016/j.ins.2024.121193.

Lin T-Y, Goyal P, Girshick R, He K, Dollar P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Litjens G, Kooi T, Ehteshami Bejnordi B, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**:60–88 DOI 10.1016/j.media.2017.07.005.

Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, Kanada K, Marinho GdO, Gallegos J, Gabriele S, Gupta V, Singh N, Natarajan V, Hofmann-Wellenhof R, Corrado GS, Peng LH, Webster DR, Ai D, Huang S, Liu Y, Dunn RC, Coz D. 2020. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* **26(6)**:900–908 DOI 10.1038/s41591-020-0842-3.

Liu Z, Jian X, Sadiq T, Shaikh ZA, Alfarraj O, Alblehai F, Tolba A. 2024. Promoted osprey optimizer: a solution for ORPD problem with electric vehicle penetration. *Scientific Reports* **14**:28052 DOI 10.1038/s41598-024-79185-6.

Liu Z, Ma Q, Zhang T, Zhao S, Gao X, Sun T, Dai Y. 2025. Quantitative modeling and uncertainty estimation for small-sample LIBS using Gaussian negative log-likelihood and Monte Carlo dropout methods. *Optics & Laser Technology* **181**:111720 DOI 10.1016/j.optlastec.2024.111720.

Lopez AR, Giro-i-Nieto X, Burdick J, Marques O. 2017. Skin lesion classification from dermoscopic images using deep learning techniques. In: *2017 13th IASTED International Conference on Biomedical Engineering (BioMed)*. Piscataway: IEEE, 49–54.

Loshchilov I, Hutter F. 2016. SGDR: stochastic gradient descent with warm restarts. ArXiv preprint DOI 10.48550/arXiv.1608.03983.

Mahbod A, Schaefer G, Ellinger I, Ecker R, Pitiot A, Wang C. 2019. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics* **71(4)**:19–29 DOI 10.1016/j.compmedimag.2018.10.007.

Mahmood T, Saba T, Rehman A, Alamri FS. 2024. Harnessing the power of radiomics and deep learning for improved breast cancer diagnosis with multiparametric breast mammography. *Expert Systems with Applications* **249**:123747 DOI 10.1016/j.eswa.2024.123747.

Marchetti MA, Liopyris K, Dusza SW, Codella NC, Gutman DA, Helba B, Kalloo A, Halpern AC, Soyer HP, Curiel-Lewandrowski C, Caffery L, Malvehy J. 2020. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the international skin imaging collaboration 2017. *Journal of the American Academy of Dermatology* **82(3)**:622–627 DOI 10.1016/j.jaad.2019.07.016.

McCaffrey C, Jahangir C, Murphy C, Burke C, Gallagher WM, Rahman A. 2024. Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer. *Expert Review of Molecular Diagnostics* **24(5)**:363–377 DOI 10.1080/14737159.2024.2346545.

Miller I, Rosic N, Stapelberg M, Hudson J, Coxon P, Furness J, Walsh J, Climstein M. 2024. Performance of commercial dermatoscopic systems that incorporate artificial intelligence for the identification of melanoma in general practice: a systematic review. *Cancers* **16(7)**:1443 DOI 10.3390/cancers16071443.

Minderer M, Gritsenko A, Stone A, Neumann M, Weissenborn D, Dosovitskiy A, Mahendran A, Arnab A, Dehghani M, Shen Z, Wang X, Zhai X, Kipf T, Houlsby N. 2022. Simple open-vocabulary object detection. In: *European Conference on Computer Vision*. Cham, Switzerland: Springer Nature, 728–755.

Mohammed JZ, Meira W Jr. 2020. *Data mining and machine learning: fundamental concepts and algorithms*. Cambridge, UK: Cambridge University.

Mostafavi Ghahfarokhi M, Asgari A, Abolnejadian M, Heydarnoori A. 2024. DistilKaggle: a distilled dataset of Kaggle Jupyter notebooks. In: *Proceedings of the 21st International Conference on Mining Software Repositories (MSR)*. Lisbon, Portugal, 647–651.

Mukhlif YA, Ramaha NT, Hameed AA, Salman M, Yon DK, Fitriyani NL, Syafrudin M, Lee SW. 2024. Ant colony and whale optimization algorithms aided by neural networks for optimum skin lesion diagnosis: a thorough review. *Mathematics* **12(7)**:1049 DOI 10.3390/math12071049.

Müller R, Duerschmidt M, Ullrich J, Knoll C, Weber S, Seitz S. 2024. Do humans and convolutional neural networks attend to similar areas during scene classification: effects of task and image type. *Applied Sciences* **14(6)**:2648 DOI 10.3390/app14062648.

Naeem A, Anees T, Khalil M, Zahra K, Naqvi RA, Lee SW. 2024. SNC_Net: skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images. *Mathematics* **12(7)**:1030 DOI 10.3390/math12071030.

Nair V, Hinton GE. 2010. Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*. Haifa, Israel, 807–814.

Naseri H, Safaei AA. 2025. Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review. *BMC Cancer* **25**:75 DOI 10.1186/s12885-024-13423-y.

Navarrete-Dechent C, Liopyris K, Marchetti MA. 2020. Multiclass artificial intelligence in dermatology-progress but still room for improvement. *The Journal of Investigative Dermatology* **141(5)**:1325–1328 DOI 10.1016/j.jid.2020.06.040.

Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, Malik P, Hluchý L. 2019. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review* **52**:77–124 DOI 10.1007/s10462-018-09679-z.

Nugroho ES, Ardiyanto I, Nugroho HA. 2023. Boosting the performance of pretrained CNN architecture on dermoscopic pigmented skin lesion classification. *Skin Research and Technology* **29(11)**:e13505 DOI 10.1111/srt.13505.

NVIDIA Corporation. 2020. The role of GPUs in accelerating deep learning and AI research. *NVIDIA Developer. Available at https://www.nvidia.com/en-us/deep-learning-ai/* (accessed 27 August 2024).

Ophir J, Cespedes I, Ponnekanti H, Yazdi Y, Li X. 1991. Elastography: a quantitative method for imaging the elasticity of biological tissues. *Ultrasonic Imaging* **13(2)**:111–134 DOI 10.1177/016173469101300201.

Ozel CB, Ozdemir HY, Dural M, Al A, Yalvac HE, Mert GO, Murat S, Cavusoglu Y. 2025. The one-minute sit-to-stand test is an alternative to the 6-minute walk test in patients with atrial fibrillation: a cross-sectional study and ROC curve analysis. *International Journal of Cardiology* **419**:132713 DOI 10.1016/j.ijcard.2024.132713.

Pang B, Nijkamp E, Wu YN. 2020. Deep learning with TensorFlow: a review. *Journal of Educational and Behavioral Statistics* **45(2)**:227–248 DOI 10.3102/1076998619872761.

Park HJ, Kim SH, Choi JY, Cha D. 2023. Human-machine cooperation meta-model for clinical diagnosis by adaptation to human expert's diagnostic characteristics. *Scientific Reports* **13(1)**:16204 DOI 10.1038/s41598-023-43291-8.

Pérez E, Ventura S. 2022. An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis. *Neural Computing and Applications* **34**:10429–10448 DOI 10.1007/s00521-021-06655-7.

Pinello K, Leite-Martins L, Gregório H, Oliveira F, Kimura KC, Dagli MLZ, de Matos A, Niza-Ribeiro J. 2025. Exploring risk factors linked to canine lymphoma: a case-control study. *Topics in Companion Animal Medicine* **65**:100948 DOI 10.1016/j.tcam.2025.100948.

Pirrera A, Giansanti D. 2023. Human-machine collaboration in diagnostics: exploring the synergy in clinical imaging with artificial intelligence. *Diagnostics* **13(13)**:2162 DOI 10.3390/diagnostics13132162.

Rahman Z, Hossain MS, Islam MR, Hasan MM, Hridhee RA. 2021. An approach for multiclass skin lesion classification based on ensemble learning. *Informatics in Medicine Unlocked* **25**:100659 DOI 10.1016/j.imu.2021.100659.

Rawat AS, Rajendran J, Sikarwar SS. 2025. Introduction to AI in disease detection—an overview of the use of AI in detecting diseases, including the benefits and limitations of the technology. In: *AI in Disease Detection: Advancements and Applications*. Piscataway: IEEE, 1–26.

Räz T. 2024. ML interpretability: simple isn't easy. *Studies in History and Philosophy of Science* **103**:159–167 DOI 10.1016/j.shpsa.2023.12.007.

Reddy DA, Roy S, Kumar S, Tripathi R. 2023. Enhanced U-Net segmentation with ensemble convolutional neural network for automated skin disease classification. *Knowledge and Information Systems* **65(10)**:4111–4156 DOI 10.1007/s10115-023-01865-y.

Reddy DA, Roy S, Kumar S, Tripathi R, Prabha N. 2024. Improved fuzzy based segmentation with hybrid classification for skin disease detection. *Procedia Computer Science* **235**:2237–2250 DOI 10.1016/j.procs.2024.04.212.

Rezvantalab A, Safigholi H, Karimijeshni S. 2018. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. ArXiv preprint DOI 10.48550/arXiv.1810.10348.

Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, 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, Reiter O, Shih G, Stratigos A, Tschandl P, Weber J, Soyer HP. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* **8**:34 DOI 10.1038/s41597-021-00815-z.

Rubegni G, Zeppieri M, Tognetti L, Cinotti E, De Piano E, D'Onghia M, Orione M, Gagliano C, Bacci T, Tarantello A, Russo NL, Castellino N, Miranda G, Cartocci A, Tosi GM, Avitabile T. 2024. Comparison of ophthalmologists versus dermatologists for the diagnosis and management of periorbital atypical pigmented skin lesions. *Journal of Clinical Medicine* **13(16)**:4787 DOI 10.3390/jcm13164787.

Sabazade S, Michalski MAL, Bartoszek J, Fili M, Holmström M, Stålhammar G. 2025. Development and validation of a deep learning algorithm for differentiation of choroidal nevi from small melanoma in fundus photographs. *Ophthalmology Science* **5(1)**:100613 DOI 10.1016/j.xops.2024.100613.

Sadeghi Z, Alizadehsani R, CIFCI MA, Kausar S, Rehman R, Mahanta P, Bora PK, Almasri A, Alkhawaldeh RS, Hussain S, Alatas B, Shoeibi A, Moosaei H, Hladík M, Nahavandi S, Pardalos PM. 2024. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering* **118**:109370 DOI 10.1016/j.compeleceng.2024.109370.

Saeed W, Shahbaz E, Maqsood Q, Ali SW, Mahnoor M. 2024. Cutaneous oncology: strategies for melanoma prevention, diagnosis, and therapy. *Cancer Control: Journal of the Moffitt Cancer Center* **31(1)**:94 DOI 10.1177/10732748241274978.

Saghir U, Singh SK, Hasan M. 2024. Skin cancer image segmentation based on midpoint analysis approach. *Journal of Imaging Informatics in Medicine* **37**:1–16 DOI 10.1007/s10278-024-01106-w.

Salman S, Liu X. 2019. Overfitting mechanism and avoidance in deep neural networks. ArXiv DOI 10.48550/arxiv.1901.06566.

Sanz-Motilva V, Martorell A, Manrique-Silva E, Terradez-Mas L, Requena C, Through V, Sanmartin O, Rodriguez-Peralto JL, Nagore E. 2024. Interobserver variability in the histopathological evaluation of melanoma: analysis of 60 cases. *Actas Dermo-Sifiliográficas* **5**: S0001-7310(24)00530-1 DOI 10.1016/j.ad.2024.05.023.

Sasseville M, Ouellet S, Rhéaume C, Sahlia M, Couture V, Després P, Paquette J, Darmon D, Bergeron F, Gagnon MP. 2025. Bias mitigation in primary health care artificial intelligence models: scoping review. *Journal of Medical Internet Research* **27**:e60269 DOI 10.2196/60269.

Schuitmaker L, Drogt J, Benders M, Jongsma K. 2025. Physicians' required competencies in AI-assisted clinical settings: a systematic review. *British Medical Bulletin* **153(1)**:ldae025 DOI 10.1093/bmb/ldae025.

Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. 2021. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making* **21**:1–23 DOI 10.1186/s12911-021-01488-9.

Selvaraj KM, Gnanagurusubbiah S, Roy RRR, Balu S. 2024. Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics. *Current Problems in Cancer* **49**:101077 DOI 10.1016/j.currproblcancer.2024.101077.

Shahid GeS, Ahmad J, Warraich CAR, Ksibi A, Alsenan S, Arshad A, Raza R, Shaikh ZA. 2025. LIU-NET: lightweight inception U-Net for efficient brain tumor segmentation from multimodal 3D MRI images. *PeerJ Computer Science* **11**:e2787 DOI 10.7717/peerj-cs.2787.

Shaikh ZA. 2018. Keyword detection techniques: a comprehensive study. *Engineering, Technology & Applied Science Research* **8(1)**:2590–2594 DOI 10.48084/etasr.1813.

Shaikh ZA. 2009. ZPD incidence development strategy for demand of ICTs in higher education institutes of Pakistan. In: *2009 Third International Symposium on Intelligent Information Technology Application*. Vol. 1. Piscataway: IEEE, 661–664.

Shaikh ZA, Datsyuk P, Baitenova LM, Belinskaja L, Ivolgina N, Rysmakhanova G, Senjyu T. 2022. Effect of the COVID-19 pandemic on renewable energy firm's profitability and capitalization. *Sustainability* **14(11)**:6870 DOI 10.3390/su14116870.

Shaikh ZA, Hajjej F, Uslu YD, Yuksel S, Dinçer H, Alroobaea R, Baqasah AM, Chinta U. 2024. A new trend in cryptographic information security for industry 5.0: a systematic review. *IEEE Access* **12**:7156–7169 DOI 10.1109/ACCESS.2024.3351485.

Shaikh ZA, Khoja SA. 2011. Teachers' skills set for personal learning environments. In: *Proceedings of the 10th European Conference on e-Learning.* Vol. 1, 762–769.

Shaikh ZA, Khoja SA. 2012. Identifying measures to foster teachers' competence for personal learning environment conceived teaching scenarios: a Delphi study. In: *Proceedings of the 13th Annual Conference on Information Technology Education*, 127–132.

Shaikh ZA, Khoja SA. 2013. Higher education in Pakistan: an ICT integration viewpoint. *International Journal of Computer Theory and Engineering* **5(3)**:410 DOI 10.7763/IJCTE.2013.V5.720.

Shaikh ZA, Khoja SA. 2014. Towards guided personal learning environments: concept, theory, and practice. In: *2014 IEEE 14th International Conference on Advanced Learning Technologies*. Piscataway: IEEE, 782–784.

Shaikh ZA, Kraikin A, Mikhaylov A, Pinter G, Hens C. 2022. Forecasting stock prices of companies producing solar panels using machine learning methods. *Complexity* **2022**:9186265 DOI 10.1155/2022/9186265.

Shaikh ZA, Laghari AA, Litvishko O, Litvishko V, Kalmykova T, Meynkhard A. 2021a. Liquid-phase deposition synthesis of ZIF-67-derived synthesis of Co3O4@ TiO2 composite for efficient electrochemical water splitting. *Metals* **11(3)**:420 DOI 10.3390/met11030420.

Shaikh ZA, Lashari IA. 2017. Blockchain technology: the new internet. *International Journal of Management Sciences and Business Research* **6(4)**:167–177.

Shaikh ZA, Moiseev N, Mikhaylov A, Yüksel S. 2021b. Facile synthesis of copper oxide-cobalt oxide/nitrogen-doped carbon (Cu_2O-Co_3O_4/CN) composite for efficient water splitting. *Applied Sciences* **11(21)**:9974 DOI 10.3390/app11219974.

Shaikh ZA, Umrani AI, Jumani AK, Laghari AA. 2019. Technology enhanced learning: a digital timeline learning system for higher educational institutes. *International Journal of Computer Science and Network Security* **19(10)**:1–5.

Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, He Z, Wong SY, Fang P, Ming WK. 2019. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Medical Informatics* **7(3)**:e10010 DOI 10.2196/10010.

Shinde S, Bigogno CM, Simmons A, Kathuria N, Ghose A, Apte V, Lapitan P, Makker S, Caglayan A, Boussios S. 2025. Precision oncology through next generation sequencing in hepatocellular carcinoma. *Heliyon* **11(3)**:e42054 DOI 10.1016/j.heliyon.2025.e42054.

Shorten C, Khoshgoftaar TM. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* **6(1)**:1–48 DOI 10.1186/s40537-019-0197-0.

Siddique SM, Tipton K, Leas B, Jepson C, Aysola J, Cohen JB, Flores E, Harhay MO, Schmidt H, Weissman GE, Fricke J, Treadwell JR, Mull NK. 2024. The impact of health care algorithms on racial and ethnic disparities: a systematic review. *Annals of Internal Medicine* **177(4)**:484–496 DOI 10.7326/M23-2960.

Smith LN. 2017. Cyclical learning rates for training neural networks. In: *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV 2017)*. Santa Rosa, CA, USA: IEEE, 464–472 DOI 10.1109/WACV.2017.58.

Soomro S, Niaz A, Choi KN. 2024. Grad++ ScoreCAM: enhancing visual explanations of deep convolutional networks using incremented gradient and score-weighted methods. *IEEE Access* **12**:61104–61112 DOI 10.1109/ACCESS.2024.3392853.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15(1)**:1929–1958.

Sterkenburg TF. 2025. Statistical learning theory and Occam's Razor: the core argument. *Minds and Machines* **35(1)**:1–28 DOI 10.1007/s11023-024-09703-y.

Stevens HP, Pellacani G, Angus C, El-Jabbour JN. 2025. Reflectance confocal microscopy to diagnose malignant melanoma and lentigo maligna in the UK: a single-centre prospective observational trial. *British Journal of Dermatology* **192(1)**:27–35 DOI 10.1093/bjd/ljae354.

Strika Z, Petkovic K, Likic R, Batenburg R. 2025. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. *Postgraduate Medical Journal* **101(1191)**:4–16 DOI 10.1093/postmj/qgae122.

Sun TS, Gao Y, Khaladkar S, Liu S, Zhao L, Kim YH, Hong SR. 2023. Designing a direct feedback loop between humans and convolutional neural networks through local explanations. *Proceedings of the ACM on Human-Computer Interaction* **7(CSCW2)**:1–32 DOI 10.1145/3610187.

Sutskever I, Martens J, Dahl G, Hinton G. 2013. On the importance of initialization and momentum in deep learning. In: *International Conference on Machine Learning*. PMLR, 1139–1147.

Tan M, Le Q. 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. PMLR, 6105–6114.

Tan L, Wu H, Xia J, Liang Y, Zhu J. 2024. Skin lesion recognition via global-local attention and dual-branch input network. *Engineering Applications of Artificial Intelligence* **127**:107385 DOI 10.1016/j.engappai.2023.107385.

Tao C, Alatas B. 2024. Real-time emotional topic recommendation in social media news using MDT and hypergraph-based neural networks. *IEEE Access* **12**:156252–156260 DOI 10.1109/ACCESS.2024.3481658.

Topol EJ. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25(1)**:44–56 DOI 10.1038/s41591-018-0300-7.

Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, Halpern A, Helba B, Hofmann-Wellenhof R, Lallas A, Lapins J, Longo C, Malvehy J, Marchetti MA, Marghoob A, Menzies S, Oakley A, Paoli J, Puig S, Rinner C, Rosendahl C, Scope A, Sinz C, Soyer HP, Thomas L, Zalaudek I, 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 DOI 10.1016/S1470-2045(19)30333-X.

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 HP, Zalaudek I, Kittler H. 2020. Human-computer collaboration for skin cancer recognition. *Nature Medicine* **26(8)**:1229–1234 DOI 10.1038/s41591-020-0942-0.

Tuba E, Bačanin N, Strumberger I, Tuba M. 2021. Convolutional neural networks hyperparameters tuning. In: *Artificial Intelligence: Theory and Applications*. Cham, Switzerland: Springer International Publishing, 65–84 DOI 10.1007/978-3-030-79347-4_4.

van den Berg J. 2024. A study on bias against women in recruitment algorithms. *Available at https://repository.tudelft.nl/file/File_58cda6ff-ec1c-4fdf-a92d-49e1c6bdf99a?preview=1*.

Wang D, Liu Y, Zhang Y, Chen Q, Han Y, Hou W, Liu C, Yu Y, Li Z, Li Z, Zhao J, Shi L, Zheng Y, Li J, Zhang R. 2024. A real-world multi-center RNA-seq benchmarking study using the Quartet and MAQC reference materials. *Nature Communications* **15(1)**:6167 DOI 10.1038/s41467-024-50420-y.

Wang G, Wong KW, Lu J. 2020. AUC-based extreme learning machines for supervised and semi-supervised imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **51(12)**:7919–7930 DOI 10.1109/TSMC.2020.2982226.

Webb AB, Berg CD, Castle PE, Crosby D, Etzioni R, Kessler LG, Menon U, Parmar M, Steele RJC, Sasieni PD. 2024. Considerations for using potential surrogate endpoints in cancer screening trials. *The Lancet Oncology* **25(5)**:e183–e192 DOI 10.1016/S1470-2045(24)00015-9.

Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, Thomas L, Lallas A, Blum A, Stolz W, Haenssle HA. 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 DOI 10.1001/jamadermatol.2019.1735.

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* **135(7)**:364–381 DOI 10.1016/j.future.2022.05.014.

Wubineh BZ, Deriba FG, Woldeyohannis MM. 2024. Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: a systematic literature review. *Urologic Oncology: Seminars and Original Investigations* **42(3)**:48–56 Elsevier DOI 10.1016/j.urolonc.2023.11.019.

Yan R, Zhou Z, Shang Z, Wang Z, Hu C, Li Y, Yang Y, Chen X, Gao RX. 2025. Knowledge driven machine learning towards interpretable intelligent prognostics and health management: review and case study. *Chinese Journal of Mechanical Engineering* **38(1)**:1–31 DOI 10.1186/s10033-024-01173-8.

Yang J, Qin H, Por LY, Shaikh ZA, Alfarraj O, Tolba A, Elghatwary M, Thwin M. 2024. Optimizing diabetic retinopathy detection with inception-V4 and dynamic version of snow leopard optimization algorithm. *Biomedical Signal Processing and Control* **96(2)**:106501 DOI 10.1016/j.bspc.2024.106501.

Yap J, Yolland W, Tschandl P. 2018. Multimodal skin lesion classification using deep learning. *Experimental Dermatology* **27(11)**:1261–1267 DOI 10.1111/exd.13777.

Yu Y, Song J, Ren Z. 2013. A new hyper-parameters selection approach for support vector machines to predict time series. In: *Pervasive Computing and the Networked World. Lecture Notes in Computer Science.* Vol. 7719. Berlin, Heidelberg, Germany: Springer DOI 10.1007/978-3-642-37015-1_68.

Zalaudek I, Argenziano G, Soyer HP, Corona R, Sera F, Blum A, Braun RP, Cabo H, Ferrara G, Kopf AW, Langford D, Menzies SW, Pellacani G, Peris K, Seidenari S, Dermoscopy Working Group. 2006. Three-point checklist of dermoscopy: an open internet study. *British Journal of Dermatology* **154(3)**:431–437 DOI 10.1111/j.1365-2133.2005.06983.x.

Zalaudek I, Kreusch J, Giacomel J, Ferrara G, Catricala C, Argenziano G. 2010. How to diagnose nonpigmented skin tumors: a review of vascular structures seen with dermoscopy. *Journal of the American Academy of Dermatology* **63(3)**:361–374 DOI 10.1016/j.jaad.2009.11.698.

Zhang W, Cheng W, Fujiwara K, Evans R, Zhu C. 2024a. Predictive modeling for hospital readmissions for patients with heart disease: an updated review from 2012–2023. *IEEE Journal of Biomedical and Health Informatics* **28(4)**:2259–2269 DOI 10.1109/JBHI.2023.3349353.

Zhang DY, Venkat A, Khasawneh H, Sali R, Zhang V, Pei Z. 2024b. Implementation of digital pathology and artificial intelligence in routine pathology practice. *Laboratory Investigation* **104(9)**:102111 DOI 10.1016/j.labinv.2024.102111.