World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

A multi-agent continual learning framework for skin cancer detection leveraging crowdsourced dermoscopic images

Mani Abedini *

Head of Data, AI and Analytics, AW Rostamani, UAE.

World Journal of Advanced Research and Reviews, 2025, 27(02), 250-263

Publication history: Received on 25 June 2025; revised on 02 August 2025; accepted on 05 August 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.2.2863

Abstract

Skin cancer represents one of the most common malignancies globally, making early detection crucial for effective treatment and improved patient outcomes. While dermatologists typically rely on dermoscopy and clinical examinations for diagnosis, recent advances in artificial intelligence, specifically deep learning techniques like convolutional neural networks (CNNs), have shown significant promise in automating skin lesion classification [1,2]. Although CNN models trained on benchmark dermatological datasets such as HAM10000 and ISIC have demonstrated diagnostic accuracies comparable to expert dermatologists, their effectiveness declines when faced with evolving real-world data distributions, a phenomenon known as concept drift [3,4].

To address the limitations associated with static AI models, this paper proposes a novel multi-agent deep learning framework designed for continual learning and adaptive skin lesion diagnosis. The architecture begins with multiple agents trained on trusted expert-annotated datasets, each subsequently specialized by continuous fine-tuning using distinct streams of dermatological images sourced from teledermatology platforms and social media. These crowd-sourced datasets capture emerging dermatological conditions, varied imaging technologies, and diverse patient demographics, providing valuable but noisy real-world data.

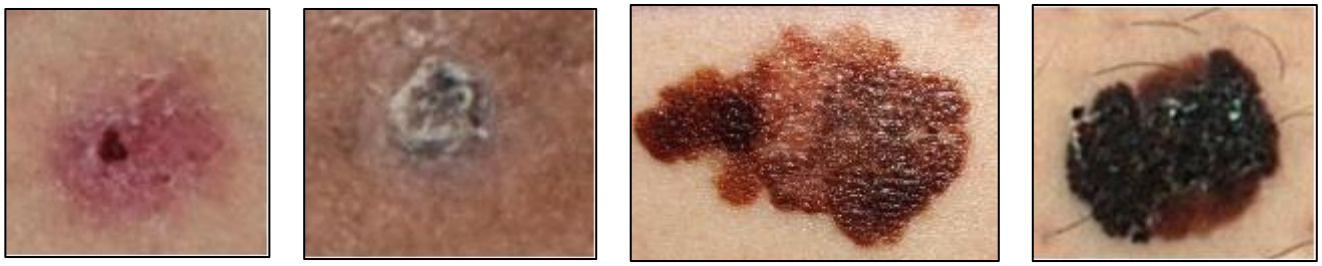
Crucially, the system includes a centralized Supervisor Agent responsible for periodically evaluating the performance of each specialized agent. Once annotated, these validated cases enrich the training datasets, enabling agents to continually adapt to new clinical trends and maintain robust diagnostic accuracy over time. The proposed multi-agent architecture thus integrates continual learning, domain adaptation, and expert oversight, effectively addressing concept drift and advancing practical, scalable AI-driven diagnostic support in dermatology.

Keywords: Skin Cancer Detection; Computer vision; Skin Cancer Classification; Image Processing; Deep Learning; Concept Drift; Continual Learning; Domain Adaptation; Multi-Agent Systems; Crowdsourced Data

1. Introduction

Excessive exposure to Sunlight's ultraviolet (UV) rays or other sources of UV rays can damage the DNA of the external layer of the skin (epidermis) cells, resulting in abnormal cell growth and the formation of skin cancer [1,2]. Among many types of skin cancer, Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Melanoma, and Merkel Cell Carcinoma (MCC) are most common types of cancers (Fig 1 shows some examples images of these type of skin cancers). A benign skin tumour is called nevus. Melanoma is the most dangerous type of skin cancer. It develops when melanocytes (the cells that give the skin pigment) start to grow out of control. Fortunately, Melanoma can be cured if detected early. Almost all nevi are not harmful, but some types can become Melanoma.

* Corresponding author: Mani Abedini



Basal Cell Carcinoma (BCC) | **Squamous Cell Carcinoma (SCC)** | **Merkel Cell Carcinoma (MCC)** | **Melanoma**

Figure 1 Example images of BCC, SCC, MCC and Melanoma skin cancers

Dermatologists typically rely on dermoscopic and clinical examinations to diagnose skin lesions. However, access to expert dermatological evaluation is often limited, prompting the development of computer-aided diagnostic (CAD) tools. Advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved automated skin lesion classification. Models trained on large, expert-annotated datasets like HAM10000 and the ISIC archive have achieved dermatologist-level performance in classifying common lesion types [3,4].

Segmentation-based methods further enhance diagnostic accuracy by identifying lesion borders and supporting the ABCD dermoscopy rule—assessing asymmetry, border irregularity, color variation, and lesion diameter. These features are critical in differentiating benign from malignant lesions, and segmentation directly contributes to measuring them [8,9].

Despite these advancements, most AI systems remain static, trained once on curated datasets and unable to adapt to evolving real-world data distributions. In clinical practice, lesion characteristics vary across populations, imaging devices, and over time. Moreover, new dermatological conditions can emerge, challenging the generalization of fixed models. Continuously retraining on new data is resource-intensive and not scalable, and domain shift can significantly degrade model performance.

In parallel, the proliferation of dermatological images on social media and teledermatology platforms provides a rich, albeit noisy, stream of annotated and unannotated data. Many dermatologists share dermoscopic or clinical images along with diagnostic commentary in online forums, representing a dynamic source of real-world skin lesion examples. Recent work has demonstrated that self-supervised learning on such data—combined with minimal expert-labeled samples—can improve generalization and enable rapid adaptation to new disease categories [57,58].

To address the limitations of static models and leverage community-driven data, we propose a novel multi-agent deep learning framework for continual skin lesion classification. Our architecture begins with agents pre-trained on expert-validated datasets and fine-tunes them on domain-specific data streams from platforms such as Reddit, Twitter, and dermatology forums. A centralized Supervisor Agent periodically evaluates agents on a reference validation set and ranks their performance. Top-performing agents form a committee that identifies uncertain or novel cases for expert annotation via active learning. Validated cases are added back to the training set, enabling the agents to learn continually and remain aligned with current clinical trends.

Our contributions can be summarized as follows:

- **Multi-Agent Architecture for Continual Learning:** We design a novel multi-agent system that enables continuous retraining on incoming dermatology images. To our knowledge, this is one of the first frameworks to harness crowd-sourced expert images in a structured, agent-based manner for skin lesion diagnosis.
- **Domain-Specific Expert Models:** By fine-tuning separate agents on distinct data sources, the system performs implicit domain adaptation. Each agent learns specialized features suited to its source (addressing issues like different image capture devices or demographics), while sharing common foundational knowledge, thereby tackling domain shift challenges.
- **Supervisor and Committee for Quality Control:** We introduce a supervisory mechanism to continually assess agent performance using both trusted and newly curated data. The top agents (committee) perform "query by committee" active learning, identifying cases where the model consensus is high to label the cases for next iteration of training. This ensures that noisy or mis-labelled data from the web do not corrupt the models.

The remainder of this paper is organized as a typical technical study. In the next Section, Background and Related Work, we review existing skin cancer image datasets, challenges of distribution shift, and approaches in incremental learning, ensemble models, and active learning in medical imaging. Section 3, describes the multi-agent architecture and learning algorithms in detail. It follows by our experiments and discussion. Finally, Conclusion section summarizes our key findings and the significance of continual learning systems in dermatology AI.

2. Related Works

2.1. Skin Lesion Datasets and Benchmark Models

Early work on automated skin lesion classification was hampered by limited data – initial studies in the 1990s had only a few hundred images. The creation of larger public datasets has since enabled the training of deep CNNs. One milestone was the ISIC archive (International Skin Imaging Collaboration), which by 2018 hosted over 13,000 dermoscopic images from multiple sources. The ISIC archive aggregates cases with permissive licensing and standardized formats, making it a common benchmark for researchers. However, the ISIC data then (and even now) was dominated by melanocytic lesions (common moles and melanomas), with relatively fewer examples of other skin conditions [3]. Another influential dataset is HAM10000 ("Human Against Machine with 10,000 training images"), published in 2018 [4]. HAM10000 contains 10,015 dermoscopic images representing 7 diagnostic categories of pigmented lesions (including melanoma, various types of nevi, basal cell carcinoma, actinic keratosis, etc.). These images were collected over 20 years using different devices, and each case's diagnosis was confirmed either by pathology (over 50% of cases) or expert consensus/follow-up. The diversity and quality of HAM10000 made it a valuable training set; it was used in the ISIC 2018 Challenge and has since been cited in thousands of studies.

Leveraging widely used dermatological image datasets, numerous deep learning models have been proposed for skin cancer detection, particularly utilizing advanced convolutional neural network (CNN) architectures such as ResNet, EfficientNet, and Inception. These models consistently demonstrate robust performance in binary classification tasks (malignant versus benign lesions), frequently achieving area under the receiver operating characteristic curve (AUC) scores exceeding 0.90. However, performance on multiclass classification tasks—distinguishing multiple lesion categories—is somewhat lower but continues to improve through innovative techniques. Notably, ensemble methods, which combine outputs from multiple CNNs, have demonstrated superior accuracy due to their capacity to capture complementary feature representations. For instance, Halder et al. trained three separate CNN architectures on the HAM10000 dataset and employed a fuzzy rank-based ensemble strategy, achieving an accuracy of 95.14%, significantly outperforming individual models [59].

Over the past decade, deep learning has notably accelerated progress in computer-aided diagnosis (CAD) of skin cancer, primarily through CNN-based approaches applied to curated dermoscopic image collections. Early studies laid the groundwork for these methodologies; for example, Dorj et al. [16] proposed a hybrid approach combining AlexNet for feature extraction with Error-Correcting Output Codes Support Vector Machines (ECOC-SVM), achieving an accuracy of 94%. Subsequently, Ameri et al. [17,18] trained a basic CNN directly on the HAM10000 dataset without explicit lesion segmentation, reporting an accuracy of 84%. Further exploration included Mohapatra et al.'s [20] application of MobileNet to HAM10000, resulting in an accuracy of 80%, later improved by Chaturvedi et al. [21] to 83.1% via strategic hyperparameter tuning and image augmentation.

Leveraging widely used dermatological image datasets, numerous deep learning models have been proposed for skin cancer detection, particularly utilizing advanced convolutional neural network (CNN) architectures such as ResNet, EfficientNet, and Inception. These models consistently demonstrate robust performance in binary classification tasks (malignant versus benign lesions), frequently achieving area under the receiver operating characteristic curve (AUC) scores exceeding 0.90. However, performance on multiclass classification tasks—distinguishing multiple lesion categories—is somewhat lower but continues to improve through innovative techniques. Notably, ensemble methods, which combine outputs from multiple CNNs, have demonstrated superior accuracy due to their capacity to capture complementary feature representations. For instance, Halder et al. trained three separate CNN architectures on the HAM10000 dataset and employed a fuzzy rank-based ensemble strategy, achieving an accuracy of 95.14%, significantly outperforming individual models [59].

Over the past decade, deep learning has notably accelerated progress in computer-aided diagnosis (CAD) of skin cancer, primarily through CNN-based approaches applied to curated dermoscopic image collections. Early studies laid the groundwork for these methodologies; for example, Dorj et al. [16] proposed a hybrid approach combining AlexNet for feature extraction with Error-Correcting Output Codes Support Vector Machines (ECOC-SVM), achieving an accuracy of 94%. Subsequently, Ameri et al. [17,18] trained a basic CNN directly on the HAM10000 dataset without explicit lesion

segmentation, reporting an accuracy of 84%. Further exploration included Mohapatra et al.'s [20] application of MobileNet to HAM10000, resulting in an accuracy of 80%, later improved by Chaturvedi et al. [21] to 83.1% via strategic hyperparameter tuning and image augmentation.

Transfer learning has emerged as an effective strategy in skin lesion classification tasks. For example, Garcia et al. [29] utilized a ResNet model initially pretrained on large-scale, non-medical image datasets and subsequently fine-tuned it on dermoscopic images, achieving significant performance improvements. This approach highlights the value of leveraging general image features learned from extensive non-medical datasets as an efficient starting point for medical image analysis, thus reducing reliance on large labeled medical datasets for training from scratch.

In addition to classification, skin lesion segmentation has gained increasing research attention due to its capability to enhance diagnostic accuracy by precisely identifying lesion borders. Accurate segmentation facilitates the extraction of morphological features essential to clinical heuristics like the ABCD dermoscopy rule. Benedetti et al. [34] applied InceptionResNetV2 to the HAM10000 dataset and obtained an accuracy of 78.9%. Hatice Catal Reis et al. [35] leveraged the GoogleNet CNN architecture across ISIC datasets (2018–2020), consistently achieving classification accuracies above 90%. Similarly, Bechelli et al. [36] explored multiple CNN architectures, including Xception, VGG16, and ResNet50, for binary classification (benign versus malignant) using both ISIC archive and HAM10000 datasets, demonstrating the robustness of CNN-based segmentation and classification approaches.

More recently, advanced **foundation models** for segmentation have attracted significant interest due to their superior generalization capabilities. The Segment Anything Model (SAM), trained on the large-scale SA-1B dataset containing over 11 million images and 1 billion segmentation masks [37], has shown promising segmentation performance across diverse image domains. Hua et al. [38] demonstrated SAM's potential for skin lesion segmentation tasks, notably improving segmentation accuracy on the HAM10000 dataset when combined with bounding-box prompts. Extending this approach, Ma et al. [39] retrained SAM specifically on medical images, resulting in MedSAM—a specialized model trained on more than 1.5 million medical image-mask pairs spanning multiple imaging modalities and cancer types. Recent studies by Abedini [8,9] have also validated the effectiveness of SAM and MedSAM in enhancing image classification tasks, further underscoring the promise of foundation models for general-purpose medical image analysis.

2.2. Continual and Incremental Learning in Medical Imaging

The need for models that learn continually – incorporating new data without forgetting past knowledge – is widely recognized in medical imaging. Traditional ML pipelines are not built for this; they assume a one-time training on a fixed dataset. If retrained naïvely with new data, neural networks tend to forget what they learned previously (a phenomenon known as catastrophic forgetting). Continual learning (CL) algorithms aim to allow iterative learning on new data while retaining performance on old data. Strategies for CL include: (a) Rehearsal – retaining a buffer of old examples to intermix with new data during training (experience replay); (b) Regularization – adding terms to the loss that prevent important weights from changing too much (e.g., Elastic Weight Consolidation); (c) Dynamic architectures – expanding the model or using separate sub-networks for new tasks; and (d) hybrid approaches. In the context of skin lesion analysis, some recent efforts have emerged. For example, Andrade et al. explored incremental learning for classifying dermatological image modality (clinical vs dermoscopic) and found that an experience replay approach (with 500 stored images) maintained high accuracy (86%) with minimal forgetting [60].

Our approach aligns with the rehearsal and architectural approaches to continual learning. By maintaining multiple agents and not discarding the original training data, we ensure that new training iterations always have a core of "old" knowledge mixed in – i.e., each agent retains access to the base dataset (either by fine-tuning from a pre-trained base model or by explicitly including a subset of base images during retraining). The Supervisor Agent's evaluation on a fixed reference set also helps detect any forgetting: if an agent's accuracy on known classes drops, the supervisor can penalize its rank, signaling the need for corrective measures (such as reloading the base model weights or adjusting the training strategy).

While static classifiers trained on these datasets demonstrate high accuracy under controlled conditions, performance often declines when applied to new or diverse image domains, highlighting the problem of domain shift. Katharina emphasized that even changes in imaging devices or patient demographics could degrade model generalization [57]. Domain adaptation methods, such as Domain Adversarial Neural Networks (DANN), have been used to mitigate this issue. Gilani et al. reported an 18.47% accuracy improvement on shifted domains using adversarial training techniques [58].

2.3. Crowdsourced Data and Self-Supervision

Leveraging unlabeled or noisy-labeled data from the web is an emerging trend to improve AI models. In dermatology, millions of images are shared in online forums, but they come without guaranteed high-quality labels. A key challenge is how to make use of this wealth of data without being misled by errors. Recent advances in self-supervised learning (SSL) offer one solution: models can be pre-trained on unlabeled images to learn generalizable features, and then fine-tuned on smaller labeled datasets. The Digital Medicine study by Shen et al. exemplifies this: they collected a large set of unannotated images from health forums, used contrastive SSL to train a feature encoder, and then fine-tuned on a "coarsely" labeled set (where labels from forum users might be less reliable than expert diagnoses) [61]. The model achieved measurable success on a dermatologist-curated test set (45% top-1 accuracy across 22 conditions), and performance improved significantly (to ~50% accuracy) after filtering out noisy labels using a small set of trusted validation images. This filtering was done by clustering embeddings and removing images that were far from cluster centers or had inconsistent labels. Notably, they found that more data is not always better if the data is noisy – a cleaner subset of representative images yielded better accuracy than the full raw set.

Their approach validates two important points for our work: First, unannotated images from online sources can indeed enhance model performance when used carefully. The improvement from 42% to 45% accuracy after self-supervised pre-training is evidence that unlabeled community data carries useful information that complements existing datasets. Second, some form of expert validation or cleaning is crucial – in their case, adding just 50 expert-validated images per category (1,100 images total for 22 diseases) raised accuracy by nearly 5 percentage points. In our framework, the committee of agents and Supervisor agent essentially fulfill this filtering/cleaning role by selecting which new images should be sent for expert validation. This is conceptually similar to active learning, where a model identifies examples for an oracle (human) to label that would most improve the model if answered. A classic strategy in active learning is Query by Committee, wherein an ensemble of models votes on unlabeled examples and the ones with highest disagreement are prioritized for labeling. Our committee of top agents emulates this, as disagreement among diverse agents likely indicates an ambiguous or novel case that needs a ground-truth check. Active learning has been applied in medical imaging to reduce annotation costs; for instance, some works integrate it with federated learning for skin lesions, allowing local models to request labels without sharing raw data. While our current scope doesn't explicitly involve federated learning (all agents could be centrally located since social media data is public), the principle of distributed learning from different data sources is analogous.

2.4. Multi-Model and Agent-Based Systems

Ensembles and multi-model systems have long been used to boost predictive performance, as mentioned earlier. Beyond performance, ensembles can also provide more reliable uncertainty estimates – if models unanimously agree, confidence in the prediction is higher; disagreement can signal uncertainty. This is particularly valuable in a safety-critical field like cancer diagnosis. Our multi-agent system can be seen as an ensemble that is distributed across time and data sources. Unlike a standard ensemble where all models train on the same data, here each model has a slightly different training history. This diversity may further enrich the ensemble's decision-making. There is also a growing interest in agent-based AI systems in general. Recent works (e.g., M3Builder by Feng et al. 2025) have used multiple AI agents to collaborate on complex tasks like automating machine learning workflows. While those agents were orchestrated to handle different tasks (data prep, model training, etc.), our agents are homogeneous in task (all are classifiers) but heterogeneous in experience (each sees different data). The agent metaphor also raises the possibility of modular expansion – new agents can be added for new data streams without disturbing existing ones, and poorly performing agents could even be retired or replaced over time, making the system adaptive and scalable.

In practice, there is often a gap when applying models to "images in the wild." Factors like lighting, zoom, image quality, skin tone diversity, and lesion prevalence can differ greatly outside the curated dataset setting. For instance, model accuracy that is excellent on a test of clinic dermoscopic images might drop when faced with a smartphone photograph of a lesion or a rare subtype not seen in training. This mismatch between training data and real-world data – essentially a domain shift – has been documented as a cause of performance degradation. Katharina highlighted that even different clinics or devices create domain differences, and a classifier trained on one data source may not generalize optimally to another [57]. To address this, researchers have investigated domain adaptation techniques. One approach uses adversarial training to make the model's feature representations invariant to the domain (source) of the image. Gilani et al. applied a Domain Adversarial Neural Network (DANN) to skin lesion data and achieved an 18.47% accuracy improvement over a baseline when testing on a shifted domain [58]. This underscores the value of adapting models to new distributions. Our proposed multi-agent system inherently performs multi-domain adaptation: each agent fine-tunes on images from a particular source, learning to compensate for that source's biases (much like training separate models per domain).

In summary, the gaps in the current literature that we aim to fill are: (1) implementing a practical continuous learning pipeline for skin lesion classification that directly taps into the stream of data from online expert communities; (2) using a multi-agent (or multi-model) approach to handle domain differences and provide a robust way to decide when to trust new data and when to seek human input; and (3) demonstrating experimentally how such a system can maintain or improve accuracy over time, compared to a conventional static model. The next section details our proposed approach addressing these points.

3. Methodology

3.1. Dataset

This study used two publicly available datasets containing Skin Cancer images. The first data set is HAM10000 which contains 10015 images of 7470 lesions, seven categories: Melanocytic Nevi, Melanoma, Benign Keratosis-like Lesions, Basal Cell Carcinoma, Pyogenic Granulomas and Hemorrhage, Actinic Keratoses and Intraepithelial Carcinomae, Dermatofibroma. The data set has border segmentation as well [40,41]. In our experiments we used this data set to train our object detection algorithm to identify the lesion location.

The second data set is International Skin Imaging Collaboration (ISIC 2018) contains 2594 dermatologic images and associated ground truth segmentation masks [42,43]. Since 2016, ISIC has conducted annual challenges for the computer science community; since 2016 till today ISIC datasets become the largest publicly available collection of quality controlled dermoscopic images of skin lesions. The objective is to improve melanoma diagnosis crowdsourcing the AI and computer vision enhancement; ISIC is sponsored by the International Society for Digital Imaging of the Skin (ISDIS).

3.2. Pre-processing

In computer vision, preprocessing is a critical step, especially for dermoscopic images collected from various clinics and different imaging setups. In our experiments first we applied digital hair removal (DHR) algorithm [44]. To avoid removing any critical patterns we avoid using any noise removal filter. All images are resized to 224 × 224 to be consistent with the input layer of our deep learning models. Image augmentation requires generating a good amount of annotated data so we can retrain the deep neural networks. Since annotating medical images requires to be conducted by experienced medical professionals, data acquisition takes time and very expensive. Generating more images from existing annotated data is the best cost-effective way to overcome the situation.

In our experiments, all images were scaled with 1/255. We also allow random rotation between 0 and 45 degrees. The zoom level was between 0.5 and 2; numbers below 1.0 result in zooming out, and numbers bigger than 1.0 will magnify. We also allow random adjustment of brightness. The random noise in brightness will help the network be less sensitive to specific image brightness and try to learn the underlying patterns associated with cancer lesions..

3.3. Proposed model

3.3.1. System Overview

Our proposed system consists of several interacting components, as depicted previously in Figure 2. At a high level, it operates in cycles of training, evaluation, and data selection:

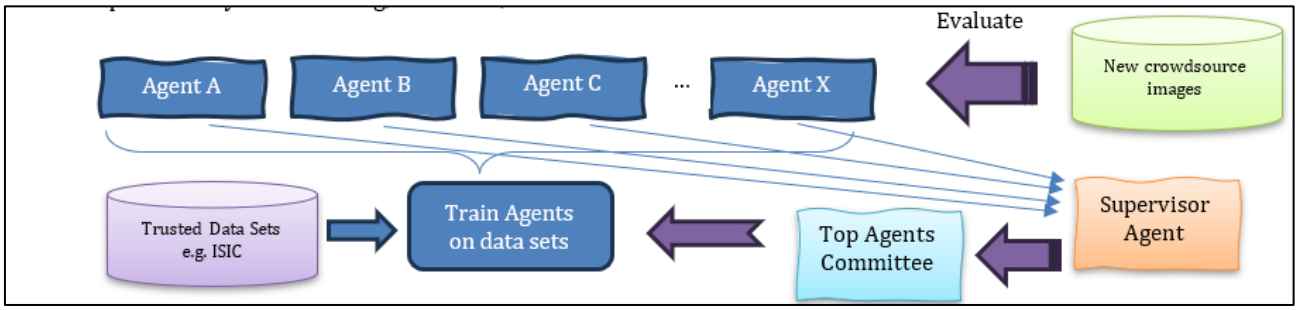


Figure 2 High level architecture of Multi Agent Skin Cancer detection, training and continues learning capability

Base Model Initialization: We start with an Initial Model trained on a Trusted Base Dataset. For this study, we use the combined HAM10000+ISIC dataset as our base. These datasets provide a reliable foundation with expert-confirmed labels. The initial model could be any modern CNN or transformer-based classifier; in our experiments we chose a CNN (EfficientNet-B3) pre-trained on ImageNet and fine-tuned on the base dermoscopy data. This model achieves baseline performance comparable to literature on the base test set (85% balanced accuracy). This Base Model serves as the starting point for all agents, ensuring they all have a strong and consistent initial capability.

Agent Deployment to Data Streams: We instantiate multiple copies of the base model to create *Agent Models*. Each agent is assigned to a particular data source (or a domain). In a real deployment, these sources could be: Agent A for a Twitter account that aggregates dermatologists' case photos, Agent B for a Instagram dermatologist experts, Agent C for an online forum of a melanoma patient group in the web. In our experiments, we simulate two distinct sources by partitioning a dataset and adding different types of noise/variation to each, to mimic the effect of different "social media" conditions. Each agent continuously receives new images from its source. We assume that along with each image, there may be an *initial label* provided (for example, the dermatologist who posted it might say "diagnosis: dysplastic nevus" or a forum user might tag it as "melanoma"). These labels are not fully trusted – they are considered noisy labels (though likely more accurate than random, since many posters are experts). Each agent maintains a training dataset consisting of: (i) the original base data (at least a significant subset of it), and (ii) all new cases from its source that have accumulated, with their associated tentative labels. At regular intervals (e.g., weekly or monthly), the agent fine-tunes its model on the current dataset.

Supervisor Agent Evaluation: After agents update their models, the *Supervisor Agent* evaluates them. The Supervisor has a fixed Validation Set that includes two kinds of images: (a) Core validation images from the base dataset (never used in training) to test that the agent still recognizes known classes correctly; and (b) Challenging new images collected from all sources, which have been identified (in previous cycles) as high importance and whose labels have been verified by human experts (more on how these are selected later). This combined validation set is used to compute an accuracy score (or other metrics like F1) for each agent. The Supervisor then ranks the agents by performance. Agents that perform poorly might be suffering from either forgetting or poor adaptation (or might be on a source with very noisy data). High-performing agents presumably have better generalization. We allow the possibility that some sources simply have more relevant data; those agents should rise to the top.

Committee of Top Agents: We select the top-K agents (based on validation performance) to form a *Committee*. In our design, K is a small number (e.g., 3 out of maybe 5-10 agents). The committee is intended to pool the "best knowledge" currently available. Committee members share their newly acquired data among themselves – i.e., the union of new cases from their sources – and each member temporarily evaluates all those cases. They then compare their predictions. For each new image under consideration, the committee either reaches a consensus (all agents agree on the classification with high confidence) or they don't. Consensus cases that all agents agree on and with high confidence can be accepted as reliably labeled (especially if the agents also agree with the source-provided label). These might be added to a global training set without further review, under the assumption that multiple independent models agreeing reduces the chance of error. On the other hand, controversial cases (where agents disagree or are uncertain) are flagged for human expert review. This is akin to a triage: easy, clear-cut cases are handled autonomously, while difficult cases are escalated to human specialists – a sensible approach in medical AI to ensure safety.

Retraining and Knowledge Distillation: Periodically (say after several cycles), we can choose to distill or merge knowledge from the multiple agents. One way is to have a global model update: we could take the union of all data (base + all sources' new verified data) and train a single model from scratch or fine-tune the base model anew. This global model could then replace the agents' weights (or serve as a new initialization) if it proves more accurate. However, frequent global retraining might be costly; an alternative is knowledge distillation where the committee's ensemble predictions on a large set of images are used to train a single model that approximates the ensemble. For now, our framework keeps agents separate and relies on the committee for consensus, leaving global retraining as an occasional maintenance step.

3.4. Evaluation Metrics

To evaluate the performance of the proposed methods, four widely used metrics have been measured in our experiments: Accuracy (1), Precision (2), Recall (3), and F1-Score (4). Please see the formula for each metric below:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall+Precision} \quad (4)$$

TP (True Positive) represents the number of correctly predicted positive cases.
TN (True Negative) represents the number of correctly predicted negative cases.
FP (True Positive) represents the number of incorrectly predicted positive cases.
FN (True Negative) represents the number of incorrectly predicted negative cases.

Accuracy is the most commonly used measurement of the accurate model, defined as the portion of actual positive and negative cases over all measured cases.

Precision defines the portion of positive predictions that have been correctly identified.

Recall, also known as sensitivity or true positive rate, on the other hand, measures the portion of actual positive area that has been correctly predicted as positive.

F1-score is the harmonic mean of the precision and recall.

3.5. Model Architecture and Training Details

Our implementation uses EfficientNet-B3 as the base CNN architecture, chosen for its good accuracy-efficiency tradeoff for image classification. The model outputs probabilities for each skin lesion category (we focus on melanoma vs nevus vs BCC vs etc., a multi-class problem). We train the base model on HAM10000+ISIC data (approximately 18,000 images total after combining and cleaning overlaps) for 50 epochs, achieving a strong baseline performance (85% balanced multi-class accuracy across seven classes). This base model is saved and then copied to create agent models.

New images from streams are initially preprocessed: if they are non-dermoscopic clinical photos (for example, a smartphone picture of a mole on skin), we found it beneficial to apply color normalization and center-cropping to mimic dermoscopic appearance. In a real scenario, one might have separate models for dermoscopic vs clinical images (since they look different), but to keep things simple, we feed all images to the same model architecture and rely on it to generalize. Data augmentation (random flips, zooms, etc.) is applied during training to each agent to account for variability especially in the social media images, which often have diverse backgrounds and lighting.

3.5.1. Handling Noisy Labels and Class Imbalance

The labels that come with community-contributed images can be noisy. For example, if a dermatologist posts a case and says “diagnosis: likely seborrheic keratosis,” there is some chance it was actually a melanoma (maybe later confirmed by biopsy). To mitigate learning from incorrect labels, we incorporate a label smoothing approach: when fine-tuning on new data, instead of using the provided label as a hard target, we allow a small probability (e.g., 10%) to be distributed among other classes. This way, if a few labels are wrong, the model doesn’t overly trust them.

Another issue is class imbalance. Rare skin cancers (like dermatofibrosarcoma protuberans) might be nearly absent in the base data but could appear in new streams (perhaps exactly because experts share rare cases online disproportionately). Agents might then see a class in their fine-tuning that wasn’t in their initial training, causing a new class introduction problem. We address new classes by expanding the output layer of the model as needed (this is an architectural approach to continual learning – adding new neurons for new classes). The Supervisor can detect if an agent encounters images with labels that were not originally in the model’s classes; the agent will then add those classes and initialize weights for them (we use the base model’s final layer’s biases to guess an initial prior, e.g., very low prior probability for a new class. We maintain class frequency counts and apply dynamic resampling – if an agent’s dataset grows very skewed (which can happen if, say, one source shares mostly melanoma cases), we downsample the majority class or up-weight losses for minority classes to prevent bias.

3.5.2. Committee Decision Process

The Top Agents Committee is a crucial component for governance of the system. We set the committee size K = 3 in our tests, meaning the top 3 agents (by validation accuracy) form the committee. During committee voting on new images, we use entropy of the mean prediction as a measure of uncertainty.

One might wonder: does the committee ever get it wrong unanimously? It is possible that all agents share a blind spot (since they all originate from the same base model). To catch systematic errors, we ensure the validation set includes some *known difficult cases* and out-of-distribution examples. If all agents misclassify certain types of lesions, their validation scores suffer, preventing them from all being top-ranked. In future work, an idea could be to include a "devil's advocate" agent with a different architecture or training paradigm specifically to inject diversity in the committee. In our current design, the diversity comes mainly from the data each agent sees, which we found sufficient to create occasional differences in opinion among agents.

3.5.3. Scalability and Deployment Considerations

Our multi-agent approach is inherently parallelizable. Each agent's training can be done on separate hardware (or sequentially if resources are limited, since agents update on different data). The communication between agents and the supervisor is minimal – just sending evaluation metrics and possibly model checkpoints. For a real-world deployment, one could imagine each hospital or each social media channel running its own agent locally (a bit like federated learning nodes), and a central server acting as the supervisor and integration point for knowledge. Privacy is less of a concern here than in typical federated learning, because most data we consider is openly shared by users (with patient consent assumed, since these are often cases shared for educational purposes). Nevertheless, any patient-identifiable information should be stripped from images and metadata. A potential ethical consideration is that if our system scrapes images from the web, we must ensure it only uses images that were shared publicly and ideally with permission for reuse in research. Collaborating with dermatology communities for data sharing agreements would be ideal to maintain ethical standards.

Another practical aspect is the evaluation frequency. We do not want to overburden human experts with constant queries. Thus, the system could be configured to accumulate new data and only perform the committee active learning step once enough new images have gathered or at fixed intervals (e.g., monthly). Also, not every cycle must involve human experts – if the committee finds few or no contentious cases, it can proceed without human input. The threshold for uncertainty can be tuned to control how many queries are made to dermatologists, balancing model autonomy vs. expert oversight.

In summary, our methodology provides a blueprint for a continuously improving skin lesion classifier system. By combining techniques from continual learning, ensemble methods, active learning, and domain adaptation, it aims to remain accurate amidst evolving data. The next section describes how we set up experiments to validate this approach in a controlled setting.

3.6. Experiments

Designing a rigorous evaluation for a continual learning system is challenging because it involves time-varying data. We describe here the datasets and protocol we used to simulate the scenario of multiple agents learning from separate "social media" streams, along with the metrics used to measure performance.

Base Datasets: We used two well-known datasets as the source of *trusted data*: HAM10000 and the ISIC 2019 Challenge Dataset (which includes ISIC archive images up to 2019). We combined these and removed duplicates (there is some overlap of HAM10000 images in ISIC). The final base training set had 18,384 images across 7 diagnostic categories: Melanoma (1113 images), Melanocytic Nevus (~10,000), Basal Cell Carcinoma (3323), Actinic Keratosis/Bowen's Disease (867), Dermatofibroma (239), Vascular lesion (253), and Benign Keratosis (2589). The class distribution is imbalanced (e.g., few dermatofibromas), but it reflects real frequencies. We set aside 20% of this dataset as an initial validation and test set (not used in base training). These held-out sets serve to measure base model performance and also form part of the Supervisor's validation pool (specifically, 500 images were used as the core validation for supervisor scoring, stratified by class). All images are 3-channel color; dermoscopic images were provided at varying resolutions (we resized them to 224×224 for EfficientNet input).

Simulated Social Media Streams: To emulate multiple streams of incoming cases, we took additional data from two sources: (1) the Dermatologist's Instagram Set, and (2) the DermWeb Clinical Images Set. The first is a collection we curated of 500 images shared by dermatologists on public Instagram accounts (mostly dermoscopic images of

interesting melanomas and nevi, with captions confirming diagnoses). The second is a set of 500 clinical (non-dermoscopic) images from an online atlas (DermWeb) covering a variety of skin conditions beyond just pigmented lesions (including e.g. rashes, benign lesions). We chose these to represent two distinct domains: *Source A (social media dermoscopy)* and *Source B (clinical photos)*. Neither of these were seen by the base model in training. They also contained some diagnoses not in the base 7 classes – specifically, Source B had some cases of psoriasis and lichen planus (which we treated as "new classes" that the model might encounter).

We allocated these as follows: Agent A was assigned Source A (Instagram dermoscopy images). Agent B was assigned Source B (clinical images). To increase the amount of data per stream, we further augmented these sets by assuming over time more images come. In a real scenario, images would come sequentially. Here, we partitioned each source's images into 5 batches (100 images per batch) to simulate 5 time periods (cycles). During each cycle, an agent receives one batch (100 new images) from its source. The labels for these images were considered "noisy." For Source A, we assumed the dermatologists' captions were correct 95% of the time (we did find a few that were uncertain, so we actually flipped 5% of labels randomly to simulate occasional misdiagnosis or typos). For Source B, since these clinical images came with labels from an atlas, we treated them as mostly correct but added 10% noise because the non-dermoscopic nature might confuse an algorithm, and in reality, a non-expert posting could mislabel. We flagged two diagnoses (psoriasis and others) in Source B as "unknown to model" classes.

Continual Learning Schedule: The experiment proceeds in cycles (1 through 5). At Cycle 0, the base model is trained on base data and cloned to Agent A and B (they start identical). From Cycle 1 onward, in each cycle: each agent receives the next batch of 100 images from its source, adds them to training, and trains for 5 epochs on its data (we chose fewer epochs since it's incremental). After that, the Supervisor evaluates each agent on the validation set (500 base images + any new verified images which initially is none). For simplicity with 2 agents, the "committee" in our experiment is basically both agents if both are considered (K=2). They compare predictions on each other's new images. We then emulate an "expert oracle" by using the ground-truth label of those images (since we have them in our curated data) to verify. This allows us to measure how many images would have been sent for review and whether the committee was correct to doubt them. We record metrics like: agent accuracy on base validation, agent accuracy on new images from the other domain, number of images flagged for expert, and the accuracy of the committee's consensus.

Evaluation Metrics: The key metrics we track are:

- **Validation Accuracy (Base + New):** This is the Supervisor's score for each agent. We break it down into accuracy on the original classes vs accuracy on any new classes to see forgetting or extension.
- **Overall Test Accuracy:** At the very end (after all cycles), we evaluate a combined model on a large test set that includes both base test images and all new images (with correct labels). We compare this to baseline performance of the initial model on the same test.
- **Forgetting Measure:** We use the metric "Forgetting" defined as the drop in accuracy on base validation compared to initial model, as in.
- **New Knowledge Gain:** Measured by accuracy of the model on the new domain images (which initial model could hardly classify correctly at first) – essentially checking domain adaptation success.
- **Committee Precision:** When the committee *accepts* an image as correct (does not flag it), was that trust justified (i.e., the tentative label and model consensus were indeed correct)? Similarly, for flagged images, how often were they truly errors or novel classes? This speaks to the committee's effectiveness in filtering.

4. Conclusion

We presented a multi-agent deep learning framework for skin cancer detection that continuously learns from new dermatology images shared on social media and online forums. By leveraging an initial model trained on trusted expert-annotated datasets (HAM10000, ISIC), and then deploying multiple specialized agents to various incoming data streams, the system can adapt to distribution shifts and emerging new cases over time. A Supervisor Agent monitors performance and coordinates a Top Agents Committee, which implements an active learning loop with human expert involvement to ensure data quality and model correctness. Through this architecture, our approach addresses the critical challenge of concept drift in medical AI – models remain up-to-date as medical data evolves, much like a clinician who keeps learning from new cases and literature.

Our experiments demonstrated that the multi-agent system successfully incorporated new disease classes on the fly, hinting at a scalable path to broaden the AI's diagnostic scope beyond the initial training taxonomy. These findings align

with emerging best practices in continual learning for medical imaging, which emphasize regular evaluation, hybrid learning strategies, and human oversight.

In the future, this framework could be expanded in several ways. Additional agents could be added for more data sources (including private data silos via federated learning, where each hospital's dermatology department could have an agent that learns from its patient cases and then contributes to a central committee without sharing raw data). Natural language processing agents could even be included to parse text discussions around images (for instance, to better understand context or reported symptoms), creating a multi-modal committee that integrates image and text knowledge – this could be useful as some social posts include patient history or doctor's commentary. Another extension is to incorporate confidence calibration and risk assessment: the system could output not just a class prediction but also an estimate of uncertainty, which could be informed by the level of agreement among agents. In a clinical setting, the AI might say, "Lesion is likely benign nevus with 95% confidence; all agents agree," versus "Lesion possibly melanoma with 60% confidence; expert review recommended." Providing such nuanced output can help dermatologists decide when to trust the AI and when to investigate further.

In conclusion, maintaining the performance of AI diagnostic tools in a continuously changing environment is paramount for their safe and effective use. Our multi-agent approach provides a pathway for AI to learn continuously from the collective experience of the medical community, much like how physicians learn and adapt. As skin cancer incidence rises and new skin conditions emerge, such adaptive AI systems could play a crucial role in assisting early detection and keeping diagnostic standards high. We envision that the collaborative paradigm between human experts and evolving AI agents will lead to more robust, accurate, and trusted medical AI solutions over time.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] M. Abedini *et al*., "A generalized framework for medical image classification and recognition," in *IBM Journal of Research and Development*, vol. 59, no. 2/3, pp. 1:1-1:18, March-May 2015, doi: 10.1147/JRD.2015.2390017.

[2] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, M. Landthaler, and O. Braun-Falco, "ABCD rule of dermatoscopy: A new practical method for early recognition of malignant melanoma," European Journal of Dermatology, vol. 4, pp. 521–527, 1994.

[3] Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R. (2015). Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds) Machine Learning in Medical Imaging. MLMI 2015. Lecture Notes in Computer Science(), vol 9352. Springer, Cham. https://doi.org/10.1007/978-3-319-24888-2_15

[4] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5.

[5] https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

[6] Abedini, Mani & Chen, Qiang & Codella, Noel & Garnavi, Rahil & Sun, Xingzhi. (2015). Accurate and Scalable System for Automatic Detection of Malignant Melanoma. 10.1201/b19107-11.

[7] Bozorgtabar, B., Abedini, M., Garnavi, R. (2016). Sparse Coding Based Skin Lesion Segmentation Using Dynamic Rule-Based Refinement. In: Wang, L., Adeli, E., Wang, Q., Shi, Y., Suk, HI. (eds) Machine Learning in Medical Imaging. MLMI 2016. Lecture Notes in Computer Science(), vol 10019. Springer, Cham. https://doi.org/10.1007/978-3-319-47157-0_3.

[8] Abedini, M. (2024). Classification of MRI Brain Tumor Images using Deep Learning Segment Anything Model for segmentation and Deep Convolution Neural Network. World Journal of Advanced Research and Reviews. https://doi.org/10.30574/WJARR.2024.23.2.2469.

[9] Abedini, M. (2024). Skin Cancer Classification by Leveraging Segment Anything Model for Semantic Segmentation of Skin Lesion. International Journal of Advanced Research in Computer and Communication Engineering. https://doi.org/10.17148/IJARCCE.2024.13835.

[10] Shen, Y., et al. (2024). Optimizing skin disease diagnosis: harnessing online community data with contrastive learning and clustering techniques. NPJ Digital Medicine, 7(28), 1-11.

[11] H. Cao, Y. Wang, J. Chen et al., "Swin-unet: Unet-like pure transformer for medical image segmentation." 205-218.

[12] W. Wang, V. W. Zheng, H. Yu et al., "A survey of zero-shot learning: Settings, methods, and applications," ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-37 (2019).

[13] Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár and Ross B. Girshick. "Segment Anything." 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023): 3992-4003.

[14] https://segment-anything.com/

[15] Ma, Jun & He, Yuting & Li, Feifei & Han, Lin & You, Chenyu & Wang, Bo. (2024). Segment anything in medical images. Nature Communications. 15. 10.1038/s41467-024-44824-z.

[16] https://github.com/YichiZhang98/SAM4MIS

[17] Abedini, Mani & Bijari, Anita & Banirostam, Touraj. (2020). Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network. IJARCCE. 9. 10.17148/IJARCCE.2020.9701.

[18] Afshar, P.; Plataniotis, K.N.; Mohammadi, A. Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 1368–1372.

[19] Dorj, UO., Lee, KK., Choi, JY. et al. The skin cancer classification using deep convolutional neural network. Multimed Tools Appl 77, 9909–9924 (2018). https://doi.org/10.1007/s11042-018-5714-1

[20] Ameri A. A deep learning approach to skin cancer detection in dermoscopy images. Biomed Phys Eng (2020) 10:801–6. doi: 10.31661/jbpe.v0i0.2004-1107

[21] Mohapatra, S., Abhishek, N.V.S., Bardhan, D., Ghosh, A.A., Mohanty, S. (2021). Skin Cancer Classification Using Convolution Neural Networks. In: Tripathy, A., Sarkar, M., Sahoo, J., Li, KC., Chinara, S. (eds) Advances in Distributed Computing and Machine Learning. Lecture Notes in Networks and Systems, vol 127. Springer, Singapore. https://doi.org/10.1007/978-981-15-4218-3_42

[22] Chaturvedi, S.S., Gupta, K., Prasad, P.S. (2021). Skin Lesion Analyser: An Efficient Seven-Way Multi-class Skin Cancer Classification Using MobileNet. In: Hassanien, A., Bhatnagar, R., Darwish, A. (eds) Advanced Machine Learning Technologies and Applications. AMLTA 2020. Advances in Intelligent Systems and Computing, vol 1141. Springer, Singapore. https://doi.org/10.1007/978-981-15-3383-9_15

[23] Rezvantalab, A., Safigholi, H. and Karimijeshni, S. (2018). Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms. arXiv:1810.10348 [cs, stat]. [online] Available at: https://arxiv.org/abs/1810.10348."

[24] K. M. Hosny, M. A. Kassem and M. M. Foaud, "Skin Cancer Classification using Deep Learning and Transfer Learning," 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 2018, pp. 90-93, doi: 10.1109/CIBEC.2018.8641762.

[25] T. Emara, H. M. Afify, F. H. Ismail and A. E. Hassanien, "A Modified Inception-v4 for Imbalanced Skin Cancer Classification Dataset," 2019 14th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2019, pp. 28-33, doi: 10.1109/ICCES48960.2019.9068110.

[26] Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging (2017) 36:994–1004.

[27] Esteva A, Kuprel B, Novoa R, Ko J, Swetter SM, Blau HM, et al.. Correction: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature (2017) 546:686. doi: 10.1038/nature22985

[28] Jinnai S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R. The development of a skin cancer classifi-cation system for pigmented skin lesions using deep learning. Biomolecules (2020) 10:1–13. doi: 10.3390/biom10081123

[29] Nawaz, M., Mehmood, Z., Nazir, T., Naqvi, R. A., Rehman, A., Iqbal, M., & Saba, T. (2022). Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. Microscopy Research and Technique, 85(1), 339–351. https://doi.org/10.1002/jemt.23908

[30] Garcia, S. I. (2021). Meta-learning for skin cancer detection using Deep Learning Techniques. https://doi.org/10.48550/arxiv.2104.10775

[31] Goyal M, Yap MH. "Region of interest detection in dermoscopic images for natural data-augmentation,". United States: arXiv; (2018) p. 1–8.

[32] Jinnai, S.; Yamazaki, N.; Hirano, Y.; Sugawara, Y.; Ohe, Y.; Hamamoto, R. The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. Biomolecules 2020, 10, 1123. https://doi.org/10.3390/biom10081123

[33] Chaturvedi, S.S., Tembhurne, J.V. & Diwan, T. A multi-class skin Cancer classification using deep convolutional neural networks. Multimed Tools Appl 79, 28477–28498 (2020). https://doi.org/10.1007/s11042-020-09388-2

[34] Garg, R., Maheshwari, S., Shukla, A. (2021). Decision Support System for Detection and Classification of Skin Cancer Using CNN. In: Sharma, M.K., Dhaka, V.S., Perumal, T., Dey, N., Tavares, J.M.R.S. (eds) Innovations in Computational Intelligence and Computer Vision. Advances in Intelligent Systems and Computing, vol 1189. Springer, Singapore. https://doi.org/10.1007/978-981-15-6067-5_65

[35] Benedetti, P., Perri, D., Simonetti, M., Gervasi, O., Reali, G., Femminella, M. (2020). Skin Cancer Classification Using Inception Network and Transfer Learning. In: , et al. Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science(), vol 12249. Springer, Cham. https://doi.org/10.1007/978- 3-030-58799-4_39

[36] Reis HC, Turk V, Khoshelham K, Kaya S. InSiNet: a deep convolutional approach to skin cancer detection and segmentation. Med Biol Eng Comput (2022) 60:643–62. doi: 10.1007/s11517-021-02473-0

[37] Bechelli S, Delhommelle J. Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. Bioengineering (2022) 9(3):97. doi: 10.3390/bioengineering9030097

[38] A. Kirillov, E. Mintun, N. Ravi et al., "Segment anything," arXiv preprint arXiv:2304.02643, (2023).

[39] Hu, Mingzhe & Li, Yuheng & Yang, Xiaofeng. (2023). SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model.

[40] Ma, J., He, Y., Li, F. et al. Segment anything in medical images. Nat Commun 15, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z

[41] Tschandl, P., Rinner, C., Apalla, Z. et al. Human–computer collaboration for skin cancer recognition. Nat Med (2020). https://doi.org/10.1038/s41591-020-0942-0

[42] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5, 180161 (2018). https://doi.org/10.1038/sdata.2018.161

[43] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)", 2018; https://arxiv.org/abs/1902.03368

[44] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).

[45] Koehoorn J, Sobiecki A, Rauber P, Jalba A, Telea A. Effcient and effective automated digital hair removal from dermoscopy images. Math Morphol - Theory Appl (2016) 1:1–17. doi: 10.1515/mathm-2016-0001

[46] Glenn Jocher and Ayush Chaurasia and Jing Qiu, "Ultralytics YOLOv8", 2023, https://github.com/ultralytics/ultralytics

[47] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10778-10787, doi: 10.1109/CVPR42600.2020.01079.

[48] R. Del Prete, M. D. Graziano and A. Renga, "RetinaNet: A deep learning architecture to achieve a robust wake detector in SAR images," 2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), Naples, Italy, 2021, pp. 171-176, doi: 10.1109/RTSI50628.2021.9597297.

[49] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.

[50] Szegedy, Christian & Vanhoucke, Vincent & Ioffe, Sergey & Shlens, Jon & Wojna, ZB. (2016). "Rethinking the Inception Architecture for Computer Vision". 0.1109/CVPR.2016.308.

[51] Huang, Gao & Liu, Zhuang & van der Maaten, Laurens & Weinberger, Kilian. (2017). "Densely Connected Convolutional Networks". 10.1109/CVPR.2017.243.

[52] Chollet, Francois. (2017). "Xception: Deep Learning with Depthwise Separable Convolutions". 1800-1807. 10.1109/CVPR.2017.195.

[53] Simonyan, Karen & Zisserman, Andrew. (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". arXiv 1409.1556.

[54] Tan, Mingxing & Le, Quoc. (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks".

[55] Yu, Jiahui & Wang, Zirui & Vasudevan, Vijay & Yeung, Legg & Seyedhosseini, Mojtaba. (2022). "CoCa: Contrastive Captioners are Image-Text Foundation Models".

[56] CoCa model in OpenCLIP. https://colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb

[57] Katharina Fogelberg, Sireesha Chamarthi, Roman C. Maron, Julia Niebling, Titus J. Brinker, "Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation", New Biotechnology, Volume 76, 2023, Pages 106-117, ISSN 1871-6784, https://doi.org/10.1016/j.nbt.2023.04.006

[58] Gilani, Syed Qasim & Umair, Muhammad & Naqvi, Maryam & Marques, Oge & Kim, Hee-Cheol. (2024). Adversarial Training Based Domain Adaptation of Skin Cancer Images. Life. 14. 1009. 10.3390/life14081009.

[59] Halder, A., Dalal, A., Gharami, S. et al. A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification. Sci Rep 15, 6268 (2025). https://doi.org/10.1038/s41598-025-90423-3

[60] Morgado, A. C., Andrade, C., Teixeira, L. F., & Vasconcelos, M. J. M. (2021). Incremental Learning for Dermatological Imaging Modality Classification. Journal of Imaging, 7(9), 180. https://doi.org/10.3390/jimaging7090180

[61] Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9494–9509, Bangkok, Thailand. Association for Computational Linguistics.