Original Article

Deep ensemble learning using a demographic machine learning risk stratifier for binary classification of skin lesions using dermatoscopic images

Ansh Roge[1,2]^, Patrick Ting[1,3], Andrew Chern[1,4], William Ting[1,5]^

[1]California Dermatology Care, San Ramon, CA, USA; [2]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA; [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA; [4]Department of Biomedical Engineering, University of California, Berkeley, CA, USA; [5]Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Contributions: (I) Conception and design: A Roge; (II) Administrative support: W Ting; (III) Provision of study materials or patients: A Roge, P Ting, A Chern; (IV) Collection and assembly of data: A Roge, P Ting, A Chern; (V) Data analysis and interpretation: A Roge, P Ting, A Chern; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Ansh Roge, BA. California Dermatology Care, San Ramon, 7536 Balmoral Way, San Ramon, CA 94582, USA; Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA. Email: rogeansh@gmail.com.

Background: Skin lesion classification through dermatoscopic images is the most common method for non-invasive diagnostics of dermatologic conditions. Feature extraction through deep learning (DL) based convolutional neural networks (CNNs) provides insight into differential attributes of skin lesions that may pertain to its malignancy. In this study, we sought to improve the performance of standard CNN architectures in skin lesion classification by providing a machine learning (ML)-derived risk score from patient demographic data.

Methods: We isolated 1,340 patients (n=2,200) from the HAM10000 dataset with ground-truth diagnoses of either melanoma or benign keratosis-like lesions. Images were split into train, validation, and test, with equal representation of each class in each phase. Baseline CNN performance was established by training 5 DL network architectures (N_i) with 3-fold cross-validation (CV); each of which employed leave-one-out CV and an early stopping criterion. Learning rate (LR) and weight decay (WD) were optimized to yield networks with the highest area under the receiver operating characteristic curve (AUC). For ML training, one-hot encoding was applied to demographic variables (age, sex, localization of lesion). This risk score was added as an additional feature in the final convolutional layers while training CNNs, yielding deep ensemble networks (E_i); all optimized parameters were the same as N_i.

Results: Amongst 7 ML classifiers, the random forest algorithm (M_{RF}) yielded the highest test AUC of 0.710. No significant difference was observed in test AUCs across DL networks (N_i=0.81±0.04) and ensemble networks (E_i=0.88±0.03), demonstrating network architecture did not significantly influence performance. A statistically significant increase in AUCs was observed in E_i compared to N_i (P=4.23E–3), indicating a significant contribution with the inclusion of a demographic risk score. Furthermore, activation maps generated for network visualization of test set images show higher specificity of differential features to inform network prediction in E_i. Average predictions on $D_{holdout}$ are significantly closer to true values in E_i compared to N_i.

Conclusions: The ensemble inclusion of a ML risk stratifier from demographic data may improve DL binary classification of dermatoscopic lesions.

Keywords: Deep ensemble learning; machine learning (ML); skin lesion classification; computer vision; melanoma

^ ORCID: Ansh Roge, 0009-0002-7895-3795; William Ting, 0000-0001-9364-5972.

Received: 01 May 2023; Accepted: 01 September 2023; Published online: 19 September 2023.
doi: 10.21037/jmai-23-38
View this article at: https://dx.doi.org/10.21037/jmai-23-38

Introduction

To this day, skin cancer remains as one of the most prevalent and deadly forms of cancer, worldwide (1). The onset of this widespread ailment can be significantly increased by a multitude of factors, including one's genetic predisposition (2), ultraviolet and sun exposure (3), smoking (4), age (5), and various other factors (6,7). Some of the more widespread skin cancers include squamous cell carcinoma (SCC), basal cell carcinoma (BCC), and malignant melanoma (MM).

In standard clinical practice, dermatologists inspect the superficial changes in nevi over time, using the widely used metric of reviewing asymmetry, irregular borders, multiple colors, large diameter, and evolving nature of nevi, also known as the ABCDEs (Asymmetry, Border, Color, Diameter, Evolution) of melanoma (8,9). While some experts can confidently determine the nature of a nevus by tracking superficial changes, the gold standard to confirm its identity would be through the microscopic determination of the intrusiveness of the melanocytes. Dermatopathologists prepare a microscopic slide of the lesion to establish any atypical features of the sample.

Following pathology-confirmed biopsy requires various treatment options depending on the severity and stage of the skin cancer, including surgical excision, superficial radiation therapy (10), Mohs surgery (11,12), or immunotherapy for metastasized lesions (13). While moderate-to-high efficacy rates have been demonstrated by these treatments, long-term cosmetic and phenotypic outcomes prove to be unfavorable to patients. The bottleneck of requiring a fast turnaround time for early detection and management of skin cancers is exacerbated by the limited number of dermatopathologists relative to the number of positive diagnoses. Furthermore, it is imperative to confirm the severity of nevi in the early stages in hopes of providing the least harmful and invasive outcome.

This may prove to be difficult for some, due to skin lesions being inconsistent in how they present. Benign lesions such as seborrheic keratoses (SK) can mimic SCC (14), a highly intrusive and prevalent non-melanoma skin cancer (NMSC). However, such benign lesions have been proven to adopt neoplastic activity through an aggregation of various genetic and environmental factors (15). Without constant monitoring of suspicious benign lesions, these can present with malignant features and can often go undetected, leading to many downstream complications.

Currently, biopsies prove to be the gold standard for obtaining and preparing samples for dermatopathology confirmation; these take the form of superficial shave excisional biopsies and deeper punch biopsies (16). Over the years, this method of obtaining samples has proven highly effective at collecting samples, yet it can present with some disadvantages. Biopsies are an invasive technique that often leaves a cosmetically-unappealing scar on the patient. Additionally, the nature of a biopsy depends on clinician preference and may result in the margins of a lesion not being included in a dermatopathology report. The time taken for pathologists to confirm the nature of a skin lesion can further delay treatment, potentially resulting in further malignancies. As such, non-invasive measures for skin cancer are in high demand.

The recent onset of artificial intelligence (AI) is becoming increasingly appealing as a non-invasive means of supplementary diagnostic measures. It combines the exponential increase in computational capabilities with complex algorithms to optimize learning given a dataset.

Highlight box

Key findings

- The addition of a machine learning risk score derived from demographic data as an extra feature can aid in deep learning (DL) binary lesion classification of dermatoscopic images.

What is known and what is new?

- Specialized convolutional neural network (CNN) architectures provide robust lesion classification and segmentation performance.
- Demographic data can differentially predispose some patients to skin cancer given constant environmental conditions.
- This study demonstrates that the inclusion of a risk score from demographic features may increase the performance of DL binary lesion classification challenges.

What is the implication, and what should change now?

- While standard DL models are highly accurate in classification tasks, their performance may be increased by supplying ensemble information as features in their final convolutional layers.

AI has established breakthroughs in a multitude of fields, including the automobile industry (17), consumer lifestyle (18), and finance (19), to name a few. In particular, AI serves as a cornerstone to increase efficiency within healthcare, as its applications can assist physicians in making diagnoses, and its resources be spread globally for minimal financial costs.

In particular, deep learning (DL) is a branch of AI that utilizes algorithms and statistical weighting to recursively train and output a variety of predictions. The branch of DL utilizes convolutional neural networks (CNNs) to mimic the processing techniques of the human brain: it gathers significant features from its input and assigns weights to inform a final decision.

There exists a myriad of studies that have compared various network architectures and have proven great success in identifying the nature of suspicious- and benign-appearing skin lesions via intricate computer vision algorithms. In a study by Brinker *et al.*, the performance of a ResNet50-based CNN was measured against 157 dermatologists on the classification of melanoma from the MClass-D skin cancer classification challenge. The results of this study indicated their CNN outperformed 136 out of 157 dermatologists (86.6%) in both sensitivity and specificity of classification (20). Similar results were obtained by Maron *et al.* in a multi-class skin cancer classification challenge, where CNN proved either higher or similar sensitivity and specificity relative to experienced dermatologists to an array of seven different classes of skin lesions (21).

Additionally, a study conducted by Hsu *et al.* (22) compared state-of-the-art DL models with the incorporation of their novel loss function HAC-LF for multi-class classification of skin lesions in the International Skin Imaging Collaboration (ISIC) 2019 Challenges Dataset. Utilization of their novel loss function prompted increased sensitivity metrics on minority classes amidst an imbalanced dataset. Furthermore, overall sensitivity, specificity, and accuracy loss metrics were greatly improved with HAC-LF, thus increasing the performance of DL models in the multi-class classification of skin lesions. Another study led by Zunair *et al.* highlighted the efficacy of implementing inter-class mapping to correct generalized data imbalance of multiple underrepresented skin lesions with their novel DL architecture, MelaNet. When compared to other Visual Geometry Group (VGG)-based CNNs, MelaNet outperformed the relatively standard methods achieving an AUC of over 0.81 and a specificity score of 0.92 (23).

Similarly, one particular study utilizing 71 different machine learning (ML) architectures achieved a sensitivity rating of 85% and a specificity rating of 86%. Yet despite these high accuracy rates of the model's capability to differentiate between benign nevi and MM, the predictions made lacked external confirmation and validation from trained professionals (24).

Ensemble learning (EL) is a technique that combines the outputs of multiple ML and/or DL models to improve the accuracy of predictions and can be applied in various ways in medical imaging diagnostics. Multiple models can be trained on different datasets, and their outputs can be combined to account for variations in image quality, patient population, and other factors that may affect the accuracy of the diagnosis. EL has been applied to medical imaging diagnostics with promising results. One study published in the *Journal of Medical Systems* in 2020 evaluated the performance of an ensemble model for the detection of COVID-19 from chest X-ray images. The ensemble model, which combined the outputs of multiple CNNs, achieved an accuracy of 0.96, compared to 0.85 for the best single model (25). By combining the outputs of multiple deep and ML models, EL can augment traditional DL algorithm responses in medical imaging diagnostics.

Skin cancer prevalence is multifactorial; diagnosis must take into account both superficial dermatoscopic data and demographic factors, such as age, sex, ethnicity, and location of a lesion (26). Various ML models have been implemented to stratify the risk of an individual based on such characteristic factors. Yet, to the best of our knowledge, few have used such ML risk stratifiers in conjunction with DL networks to inform the nature of a particular skin cancer lesion. The application of EL in the diagnosis of skin lesions from dermatoscopic and demographic data may prove to increase the performance of traditional DL networks.

The goal of this study was to compare the effectiveness of an adjunctive ML risk stratification model on DL skin lesion classification. We aim to determine if the addition of an ML-based risk probability as an independent channel in a DL network's decision tree will aid in the performance of skin cancer classification.

Methods

HAM10000 dataset

For this study, we utilized the HAM10000 (Human Against Machine with 10,000 training images) dataset (27),

Table 1 Distribution of lesions in HAM10000 dataset

Diagnosis	Count
Melanocytic nevi	6,705
Melanoma*	1,113
Benign-keratosis lesion*	1,099
Basal cell carcinoma	514
Actinic keratosis	327
Vascular skin lesion	142
Dermatofibroma	115

*, lesions included in this study. HAM10000, Human Against Machine with 10,000 training images Dataset.

a public dataset consisting of n=10,015 de-identified dermatoscopic images of skin lesions coming from N=7,470 patients. Images were collected from the Department of Dermatology at the Medical University of Vienna, Austria, and the skin cancer practice of Cliff Rosendahl in Queensland, Australia. Over 50% of lesions had ground-truth diagnosis confirmed by pathology, while the rest were confirmed by either subsequent follow-up, expert consensus, or *in vivo* confocal microscopy. For the purposes of this study, all data was included regardless of the method for ground-truth diagnosis.

Lesion classes included in this study were actinic keratoses, BCC, benign keratoses (a general term to include SK, lichen planus keratoses, solar lentigines, and various other benign lesions), dermatofibroma, melanocytic nevi (a general term to include dark, symmetrical, benign neoplasms of melanocytes that present in contrast to melanoma), melanoma, and vascular skin lesions (a general term to include benign blood-related growths including cherry angiomas, angiokeratomas, and pyogenic granulomas).

To report data in the setting of a binary classifier, only lesions of classes benign keratoses and melanoma (N=1,341, n=2,200) were retained for further testing due to their similar presentation and equal representation in the HAM10000 dataset. One-hot encoding assigned binary labels to benign keratoses [0] and melanoma [1]. A distribution of the number of lesions can be observed in *Table 1*.

Finally, included in the dataset was demographic information of patients such as age, sex, and lesion location. Images from the original HAM10000 dataset had segmentation masks available, but these were not used for this study.

Pre-processing

Lesions belonging to melanoma or benign keratoses was split into train (D_{train}; 49%), validation (D_{val}; 21%), and test ($D_{holdout}$; 30%), with an equal distribution of both classes in each phase (*Figure 1*).

Images were resized to a 224×224×3 pixel size and retained in traditional RGB format. Various artifacts such as hair and dead skin occluded some lesions of interest. The DullRazor algorithm (28) was implemented to remove such artifacts prior to further pre-processing. *Figure 2* depicts the clarity observed across several lesions after noise removal was performed.

Data augmentation

Isolating patients with diagnoses of either melanoma or benign keratoses resulted in a total of 2,212 lesions across 1,341 patients. Multiple images of the same lesion were included, as they were taken from different magnifications, angles, and camera qualities. To further augment the training data available, each image was subjected to vertical flip, horizontal flip, random color shifts, and random rotation (*Figure 3*). A 4-fold augmentation was performed on each training set image, increasing D_{train} to a total of 5,415 lesions.

ML algorithms

Decision tree

The decision tree algorithm in ML is among the most common, robust, and customizable models. It recursively iterates through features in the data until a criterion is met. Trees are composed of nodes and leaves, with the higher nodes indicating branch points of highly differential features. Leaves represent specific features that can be used to separate data. Seeing as users can manipulate the number of nodes and leaves present in models, this simple algorithm is among the most variable, thereby serving purposes ranging from classification to regression (29).

Gradient boosting

Gradient boosting is a common application in ML models to fine-tune hyperparameters within a model to isolate the unique combination yielding greatest performance. The promise of gradient boosting comes from its ability to aggregate several decision trees and assign weights

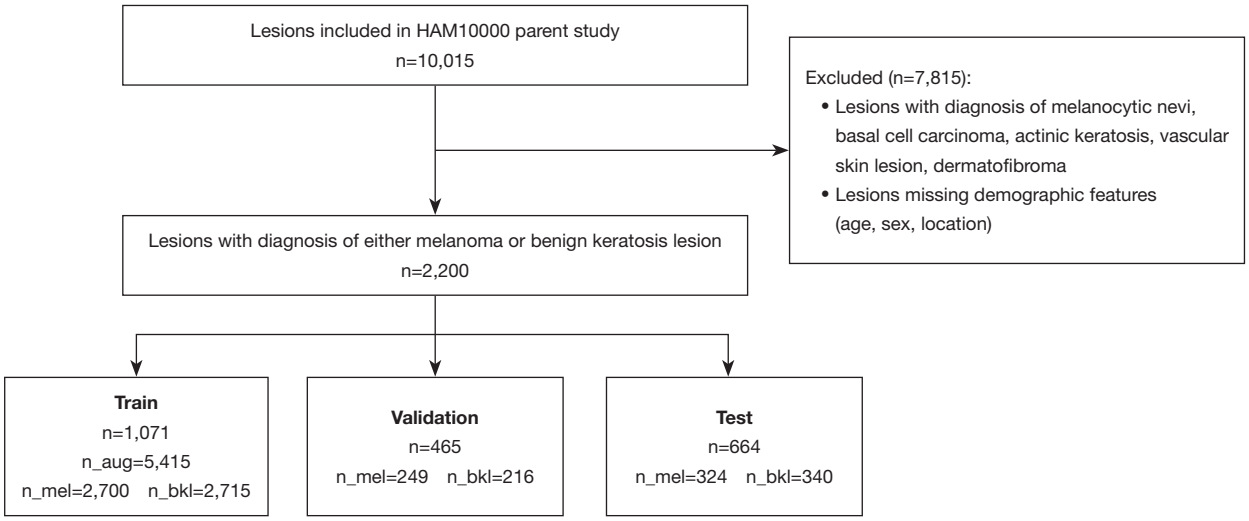

Figure 1 Patient inclusion criteria and distribution into train, validation, and test. HAM10000, Human Against Machine with 10,000 training images Dataset.

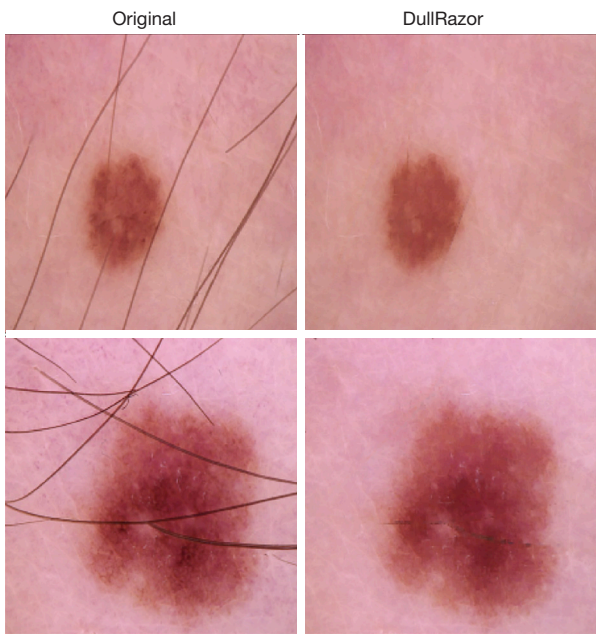

Figure 2 DullRazor algorithm effectively isolates the lesion while removing artifacts (hair, keratin debris).

to generate a composite prediction. Its performance is characterized by the performance of multiple poorly performing decision trees to create an ensemble algorithm that finetunes its weights with successive training iterations. Seeing as gradient boosting often adjusts a multitude of parameters across multiple decision trees, it is among the more computationally intensive algorithms, but has demonstrated incredible performance in data clustering and image classification challenges (30).

Naive Bayes

Naive Bayes is a common ML algorithm used in multi-categorical tasks, most prevalent being sentiment analysis in natural language processing. It assigns classes to features, assumes independence of those features, and determines an overall probability using the aggregate sum of the different classes included in the training set. While it can be informative in discrete classification tasks, this algorithm places a high importance on individual features, making it sensitive to outlier data (31).

Random forest

A random forest algorithm is a commonly used supervised learning algorithm that samples a random set of data from the training set and constructs a decision tree for each sample. A process similar to voting occurs through averaging all the individual decision trees; higher weighted values become the final prediction output. This usage of multiple different models is called EL that utilizes bagging, which is creating various training sets for each model, as well as boosting, which is combining models together to produce an even more accurate model (32).

Support vector machine (SVM)

A SVM algorithm aims to define an optimal hyperplane

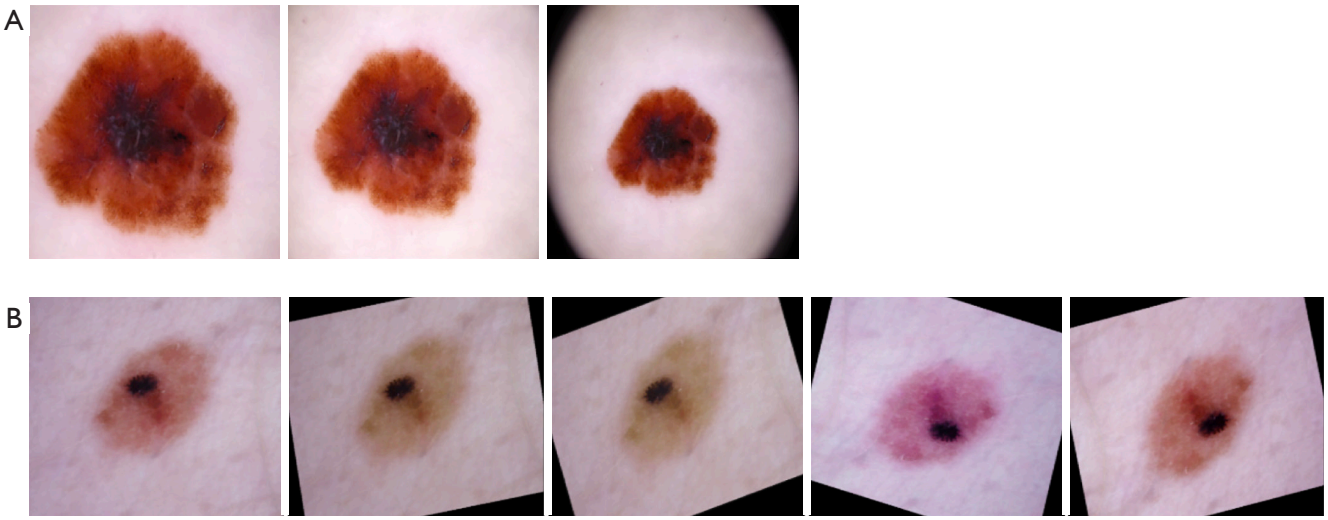

Figure 3 Example of natural and computational augmentations used to increase training data in D_{train}. (A) Natural augmentations from different camera angles. (B) Random horizontal and vertical flipping, rotations, and color shifting on all lesions in D_{train}. D_{train}, Train Dataset.

in a given space that classifies data points. Support vectors are data points with the closest proximity to the defined hyperplane. They hold the greatest weight in determining the position and orientation of the hyperplane classifier. By maximizing the distance between the support vectors and the hyperplanes, or margins, the SVM will be able to accurately distinguish between two classes of data (33).

Logistic regression

The main purpose of logistic regression algorithms is to predict the probability of data classes based on dependent variables. Using the sigmoid function, the algorithm labels certain outcomes with a probability between 0 and 1, with 1 indicating a high similarity to a compared class. The algorithm fits an equation aimed at determining a threshold to separate classes. As a result of the function's output being between 0 and 1, this method is almost exclusively used for binary classification tasks (34).

k-Nearest Neighbors (KNN)

The KNN is a supervised algorithm that excels at data classification and discriminant analysis when there is little prior knowledge of the database of interest. When plotted, categorical training data can cluster into discernable groups. Testing data, with no assigned output, can be classified based on the number of nearest neighbors to the datapoint. k is the hyperparameter that defines the number of nearest neighbors to classify a point of interest. With k=1, the unknown datapoint is grouped into the category with the nearest neighbor. Higher k allows for more neighboring datapoints to be included in the classification task. The unknown datapoint is classified by the category with the greatest amount of nearest neighbors. However, overly-elevated k can result in the over-representation of a category with few samples. KNN is an efficient algorithm for small-scale data but has a high sensitivity to outliers and large categories (35).

DL network architectures

DL methodologies can vary significantly in performance (36). While factors such as learning rate (LR), weight decay (WD), and structure of the main classifier function can yield different results, the inherent structure of DL networks plays an important role in classification performance. In particular, the recent advancement of CNNs has provided many breakthroughs in image segmentation (37) and classification challenges (38). They rely on recurrent neurons for feature extraction, proving themselves to be an efficient means for identifying heterogeneous features amongst complex datasets. Among the best-performing CNN architectures historically include AlexNet, DenseNet, MobileNetV2, ResNet, and SqueezeNet.

AlexNet

AlexNet is a CNN that preserves lightweight functionality with in-depth performance. It consists of eight layers of

trainable weights, including five convolutional layers and three fully connected layers. AlexNet uses ReLU activation functions, local response normalization, and dropout to prevent overfitting. It also employs data augmentation techniques during training to improve generalization. Initially adapted to the ImageNet challenge (39), it has the option to use pretrained weights in the training process (40).

DenseNet

DenseNet is a deep neural network architecture that is unique due to each layer being interconnected between each other. Layers are connected in a feedforward fashion, and each layer receives the feature maps of all preceding layers as input. While the vast number of connections between layers results in a larger memory size and greater length of training, it allows for feature propagation and reduces the number of parameters in the network. DenseNet has shown strong performance in a variety of image classification tasks and has been used as a base architecture for many other computer vision models (41).

MobileNetV2

MobileNetV2 is a deep neural architecture adapted for the purpose of lightweight training, initially with the intention to train on mobile devices with less computational power. Its lightweight framework utilizes depthwise separable convolutional layers, effectively separating the filtering and combining of inputs into one output with separate layers; standard convolutional layers perform this task in one layer. Multiple pretrained model weights exist using ImageNet challenge training data and are utilized for more niche image classification tasks (42).

ResNet

ResNet is among the most widely used deep neural network architectures in image classification tasks. Its unique means of optimizing backpropagation of weights enables accurate training. Additionally, its implementation of skip connections (43) allows this network architecture to pass information from earlier layers to later layers, effectively reducing the incidence of overfitting data. Use cases of this architecture have been used in a wide range of computer vision tasks such as image classification, object detection, and segmentation, with deeper variants demonstrating greater performance (44).

SqueezeNet

SqueezeNet is a lightweight neural network that boasts fifty times fewer parameters than the traditional AlexNet framework. Its lightweight architecture results in faster training time and reduced intensive computational load. One distinguishing feature is its use of fire modules, which consist of a squeeze layer and expand layer to achieve similar performance as other models without increasing model complexity (45).

Network training

Batch training was conducted on Google Colab (46) to efficiently determine optimal parameters for downstream testing. Google's free-to-use cloud GPU service granted limited access to Nvidia K80 or Nvidia T4 GPUs which were accessible through a Jupyter Notebook kernel. Model generation was conducted on a 2021 Macbook Pro with the M1 Pro chip and 16 GB of RAM. The Apple Metal Performance Shader (MPS) was the compute engine to generate the final models once the optimal parameters were established.

Statistical analysis

Confusion matrices

In binary classification, the true positive (TP) and true negative (TN) is defined as the number of positive and negative classes, respectively, correctly predicted by a model. Conversely, the false positive (FP) and false negative (FN) are the number of incorrectly attributed cases compared to the ground truth. A visual interpretation of these results can be seen in a confusion matrix.

Receiver operating characteristic (ROC) curves

A ROC curve is a visual representation of assigning a metric to the accuracy of predictive output in classification tasks. The ROC curve plots sensitivity {Eq. [1]} versus 1 – specificity {Eq. [2]}.

$$Sensitivity = \frac{TP}{TP + FN} \quad [1]$$

$$Specificity = \frac{TN}{TN + FP} \quad [2]$$

Maximal area under the ROC curve (AUC) indicates greater agreement between network predictions and ground truth values, taking the value of 1.0. In contrast, a network operating without distinguishability between classes would have an AUC of 0.5, indicating it may be operating on

random chance. AUC values ranging from 0.8–1.0 are considered to have high predictive power.

Student's *t*-test

The Student's *t*-test is a widely used metric to compute whether there exists a statistically significant difference between the means of two groups, sample 1 and sample 2. The test calculates a *t* value based on the mean (x_1, x_2), standard deviations (s_1, s_2), and sample sizes (n_1, n_2) {Eq. [3]} and compares this value to a *t* distribution. A corresponding probability, or P value (P), is calculated from the area of the *t* value in the *t* distribution and this probability determines the likelihood of the results occurring through chance.

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad [3]$$

Ethical statement

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). An ethics review board was not necessary for this project due to the nature of public datasets being used and no direct interaction with animal tissue. Patient data was confidentially stripped of identifying features in the public dataset through the institution's proprietary means.

Experimental design

Experiment 1: comparison of multiple DL architectures on skin cancer image classification performance

Five total networks (N_i: N_A, N_D, N_M, N_R, N_S) were trained in this experiment using PyTorch. All networks utilized weights from pre-trained ImageNet challenges. Solely pre-processed images from D_{train} were passed through the networks. An early-stopping criterion was implemented with leave-one-out cross-validation (CV) to prevent overfitting with a patience of 15 epochs. Models continued training until the lowest validation loss was achieved after evaluating on D_{val}, of which, the model was preserved for further comparison. Parameters such as LR, WD, and custom classifier functions were fine-tuned to save the model with the highest validation AUC, which was preserved for further downstream testing. A 3-fold CV was applied and the reported metrics were the average values of the models.

Network predictions were extracted on $D_{holdout}$ and subsequent AUCs were generated to assess network performance. Furthermore, gradient-weighted class activation maps (Grad-CAM) (47) were generated on $D_{holdout}$ to visualize differential features that networks perceived as contributing to decisions. For each model within the 3-fold CV, Grad-CAM images were generated and averaged together to generate a consensus heatmap of network importance. The complete flowchart of execution can be found in *Figure 4*.

Experiment 2: evaluating the change in performance with the addition of a ML risk stratifier on DL skin cancer image classification

Demographic data provided by the HAM10000 dataset used for this study included age of the patient, the biological sex of patients, and location of lesions. One-hot encoding was implemented on biological sex and location of lesions to numerically categorize the classes included for training (*Tables 2*,*3*). Subjects were excluded (n=12) if they had a missing value in either of the three demographic factors, following the same decision criteria represented in *Figure 1*.

ML algorithms (M_i) tested included random forest (M_{RF}), gradient boost (M_{GB}), decision tree (M_{DT}), k-Nearest Neighbor (M_{KNN}), linear regression (M_{LinReg}), logistic regression (M_{LogReg}), Gaussian Naive Bayes (M_{GNB}), multinomial Naive Bayes (M_{MNB}) and support vector machine (M_{SVM}). Various parameters specific to individual ML models were adjusted and optimized in the training phase. After pre-processing the tabular data to enumerate all values, data was split into train and test following the same splits as Experiment 1. However, the nature of ML algorithms does not require a validation set. As such, validation data was included in the training set ($D_{train, ML}$) with a holdout test set ($D_{holdout}$) for evaluation. The algorithm within M_i with the highest AUC on $D_{holdout}$ was preserved for further experimentation.

During ensemble training of DL and ML models, all parameters (LR, WD, and custom classifiers) for DL networks were kept the same as in Experiment 1. M_i was trained on the tabular demographic data first. As patient images were fed into DL training and validation networks, a predictive risk was generated by M_i given the patient's demographic data. Augmented image data in D_{train} had the same demographic features as the parent image, meaning the patient's original demographic data was used for their augmented images in the train set. The risk score was passed into the DL network as an additional feature. We modified the last convolutional layers within the feature and classifier layers of the DL networks to incorporate the extra

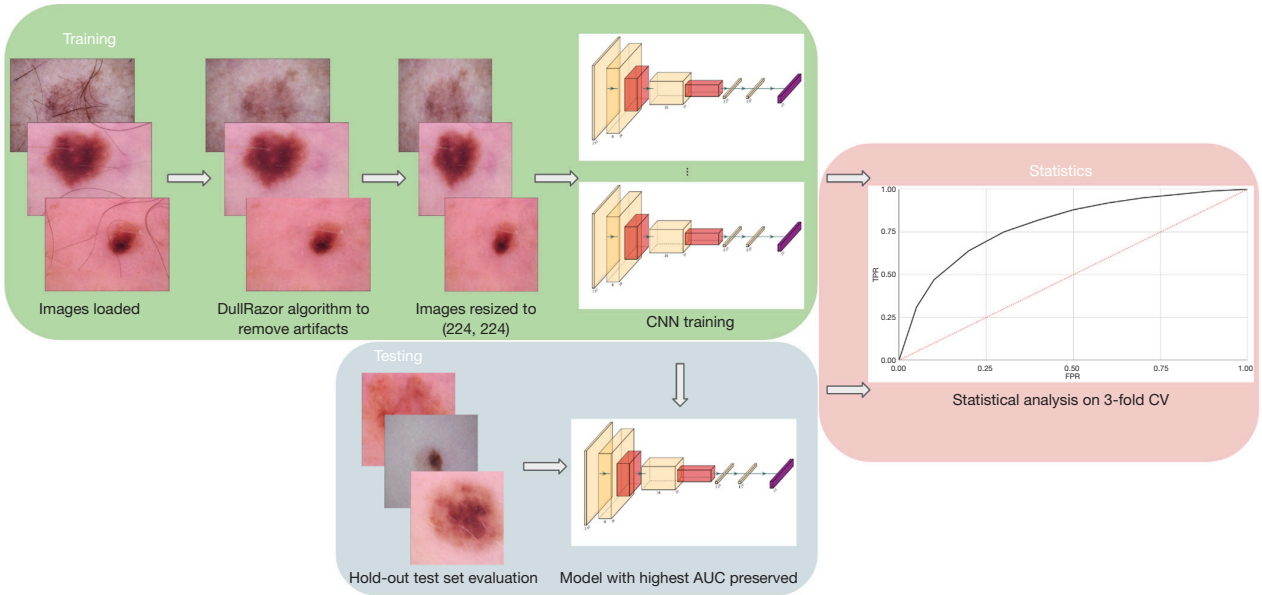

Figure 4 Pipeline of training and testing binary classification capabilities of deep networks using solely pre-processed dermatoscopic image data. CNN, convolutional neural network; AUC, area under receiver operating characteristic curve; TPR, true positive rate; FPR, false positive rate; CV, cross-validation.

Table 2 One hot encoding breakdown of sex

Sex	Numerical assignment
Male	0
Female	1

Table 3 One hot encoding breakdown of lesion location

Location	Numerical assignment
Abdomen	0
Acral	1
Back	2
Chest	3
Ear	4
Face	5
Foot	6
Genital	7
Hand	8
Lower extremity	9
Neck	10
Scalp	11
Trunk	12
Upper extremity	13

feature being passed into the networks and continued to output a singular classification value.

Subsequent training after the modification proceeded similarly to Experiment 1, where leave-one-out CV was employed to isolate the ensemble networks with the lowest validation loss. Five total EL models (ensemble referring to DL networks being informed by dermatoscopic image and ML risk stratification from demographic data) were generated via PyTorch (E_i: E_A, E_D, E_M, E_R, E_S), with each architecture receiving a 3-fold CV. Further testing was performed on $D_{holdout}$, where the ML model provided a risk stratification based on $D_{holdout}$ demographic data and was included as a feature in E_i during evaluation. Predictions were reported as the average of the 3-fold CV generated for each network architecture. Averaged Grad-CAM images were generated to highlight differential features being observed by models. The complete flowchart of execution can be seen in *Figure 5*.

Results

Experiment 1: comparison of multiple DL architectures on skin cancer image classification performance

The primary metric to denote network performance was the AUC value calculated from prediction outputs and

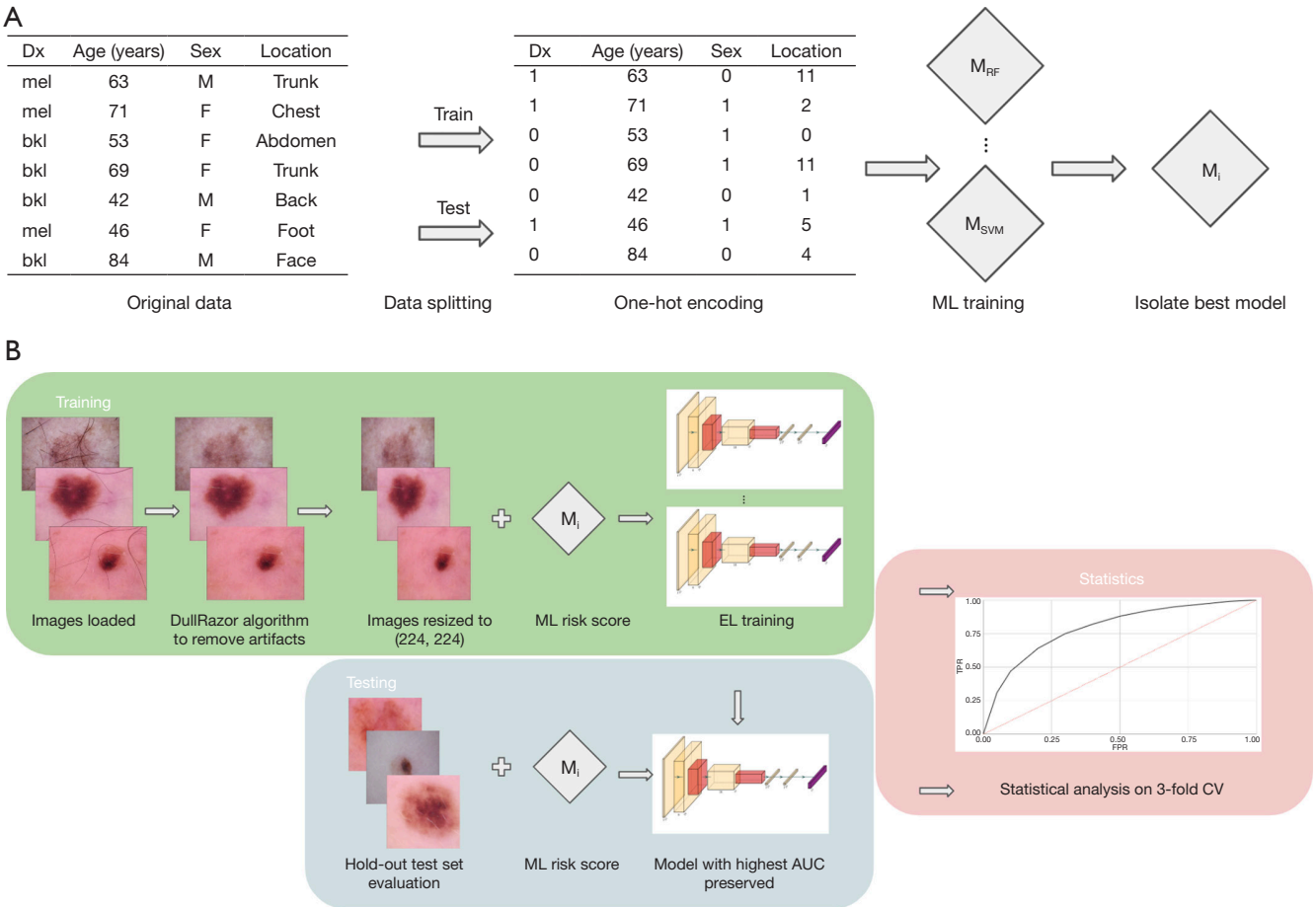

Figure 5 Pipeline of training and testing binary classification capabilities of deep ensemble networks using solely pre-processed dermatoscopic image data. (A) Training pipeline for ML models. (B) Training and testing pipeline for DL models with integration of M_i risk score. Dx, diagnosis; mel, melanoma; bkl, benign-keratosis lesion; M, male; F, female; RF, random forest; SVM, support vector machine; M_i, all ML models; ML, machine learning; EL, ensemble learning; AUC, area under receiver operating characteristic curve; TPR, true positive rate; FPR, false positive rate; CV, cross-validation; DL, deep learning.

ground-truth classification of lesions. *Table 4* demonstrates the performance of various network architectures on the predictive performance of the hold-out test set. Network AUCs throughout N_i had little variance (F=0.0, P=1.0), indicating there were no significant differences in performance across the network architectures used.

Experiment 2: evaluating the change in performance with the addition of a ML risk stratifier on DL skin cancer image classification

Out of the nine ML models trained (M_i), the random forest classifier (M_{RF}) yielded the highest test AUC of 0.710 (*Table 5*). Due to M_i being trained on three demographic variables (age of patient, biological sex, and location of lesion), lower AUCs were expected as compared to traditional datasets with an abundance of demographic features.

Preserving all parameters from Experiment 1, the risk score generated by M_{RF} on patients' demographic data was included as an additional feature when training E_i. *Table 6* depicts the performance of E_i on $D_{holdout}$. Similar to Experiment 1, AUCs across networks in E_i yielded no significant differences (F=0.0, P=1.0).

Evaluation of performances between deep and ensemble networks

The average runtime across the three-fold CV was

Table 4 Performance of network architectures in N_i with 3-fold CV on $D_{holdout}$

Architecture	Learning rate	Weight decay	AUC
AlexNet	1.00E–05	1.00E–05	0.823
DenseNet	1.00E–05	1.00E–05	0.797
MobileNetV2	1.00E–05	1.00E–03	0.822
ResNet50	1.00E–05	1.00E–02	0.772
SqueezeNet	1.00E–05	1.00E–05	0.853

N_i, all deep networks; CV, cross-validation; $D_{holdout}$, Holdout Test Dataset; AUC, area under receiver operating characteristic curve.

Table 5 Performance of ML algorithms on patient demographic data

Model type	Test AUC
Random forest	0.710
Gradient boost	0.704
Decision tree	0.670
KNN	0.621
Linear regression	0.560
Logistic regression	0.560
Gaussian naive bayes	0.551
SVM	0.547
Multinomial naive bayes	0.498

ML, machine learning; AUC, area under receiver operating characteristic curve; KNN, k-Nearest Neighbors; SVM, support vector machine.

computed for each network architecture in N_i and E_i (*Figure 6*). The endpoint was determined after models completed leave-one-out CV with a patience of 15 epochs. This analysis was conducted with the models generated using the MPS compute engine. Throughout all network architectures, E_i yielded longer runtimes (F=0.99, P=0.0029), indicating the ML prediction of demographic data and inclusion of an extra feature in deep ensemble training can significantly increase training time. The average difference in training and validation time between E_i and N_i was 16.4±6.84 minutes.

The average network performance across N_i yielded a mean test AUC of 0.8134 while the mean test AUC of E_i was 0.878. A statistically significant increase (F=0.47, P=0.0042) in performance was observed with EL for the binary classification task of skin lesions (*Figure 7*).

Table 6 Performance of network architectures in E_i with 3-fold CV on $D_{holdout}$

Architecture	Learning rate	Weight decay	AUC
AlexNet	1.00E–05	1.00E–05	0.910
DenseNet	1.00E–05	1.00E–05	0.853
MobileNetV2	1.00E–05	1.00E–03	0.881
ResNet50	1.00E–05	1.00E–02	0.872
SqueezeNet	1.00E–05	1.00E–05	0.876

E_i, all deep ensemble networks; CV, cross-validation; AUC, area under receiver operating characteristic curve; $D_{holdout}$, Holdout Test Dataset.

Upon evaluation of the performance differences between individual models, all EL networks demonstrated a significant increase in classification compared to their DL counterparts, with the exception of the SqueezeNet architecture. This architecture is known for its lightweight nature and significant performance in classification tasks. Given the notably fewer parameters required for SqueezeNet training, the relative weight of an additional ML risk feature may not have influenced the performance of the architecture as compared to other denser architectures tested. For instance, the ResNet50 architecture utilizes a particularly parameter-heavy approach with skip connections in certain training cases avoiding overfitting. This architecture demonstrated a difference of AUCs of 0.1 between N_R and E_R, the greatest difference observed between all other architectures.

Though AUCs demonstrate the classification of networks, visualizing network predictions through Grad-CAM images can provide insight into certain differential features that networks use to inform classification decisions. Grad-CAM images were generated for all networks (N_i and E_i) on $D_{holdout}$. *Figure 8* demonstrates the regions of significant features observed by each model architecture across both DL and EL modalities. Though there were inherent variations in the agreeability of differential features across architectures, networks in E_i demonstrated a higher degree of specificity and accuracy when isolating regions to inform network decision-making.

In *Figure 8A*, both modalities of the SqueezeNet architecture deliver promising activation maps, yet N_S has a broader range of highlighted features that span into seemingly healthy-appearing skin. However, when

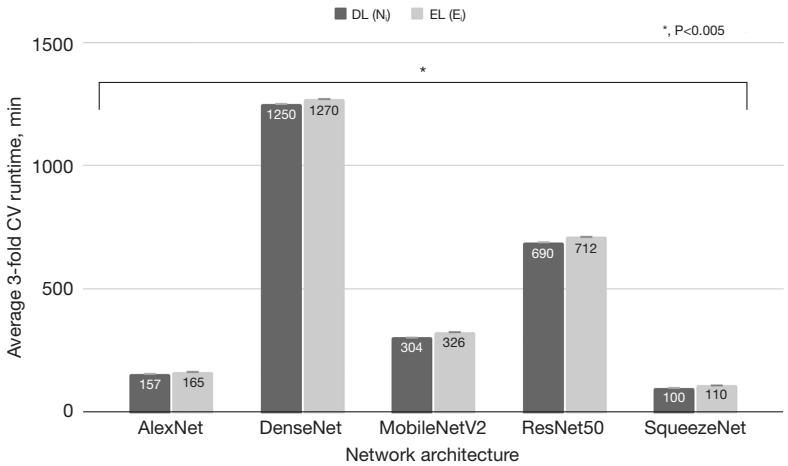

Figure 6 Average runtime analysis of N_i *vs.* E_i, shown in minutes. (*) denotes the statistically significant increase in average runtimes noted across E_i compared to N_i. DL, deep learning; EL, ensemble learning; CV, cross-validation; N_i, all deep networks; E_i, all deep ensemble networks.

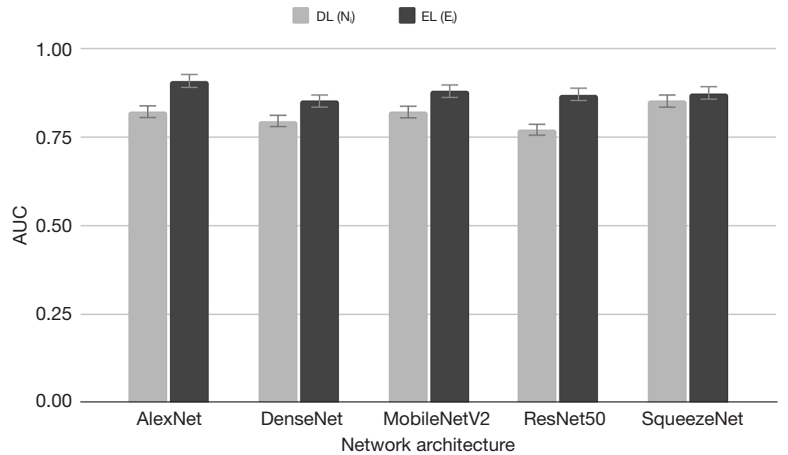

Figure 7 Comparison of AUCs across deep and ensemble methods of skin lesion classification on $D_{holdout}$. AUC, area under receiver operating characteristic curve; DL, deep learning; EL, ensemble learning; N_i, all deep networks; E_i, all deep ensemble networks; $D_{holdout}$, Holdout Test Dataset.

observing the activation map produced by E_S, the same general regions are highlighted, but the outline takes a more defined shape to the lesion. Furthermore, the primary body highlighted refers to a significantly more hyperpigmented and irregular component of the nevus. This difference indicates E_S informs its decision on differential features that are characteristic of melanoma, including the hyperpigmented nature and irregular borders of the nevus.

A different story is demonstrated in *Figure 8B*, where N_i highlights the benign lesions while sparing surrounding healthy skin. Upon analyzing the activation map generated by N_D, the most notable region highlighted on the image of the benign keratosis is the lesion itself. When looking at the activation map produced by its counterpart, E_D, regions of healthy tissue are increasingly highlighted while the lesion itself is seen as significantly benign. This pattern can be seen with all networks in E_i, where the benign lesion contains less highlighting relative to the surrounding healthy skin. The result is a more composite prediction that takes into account features present throughout the image in E_i, marking the nevus itself as benign.

In the setting of implementation of the ML risk stratifier

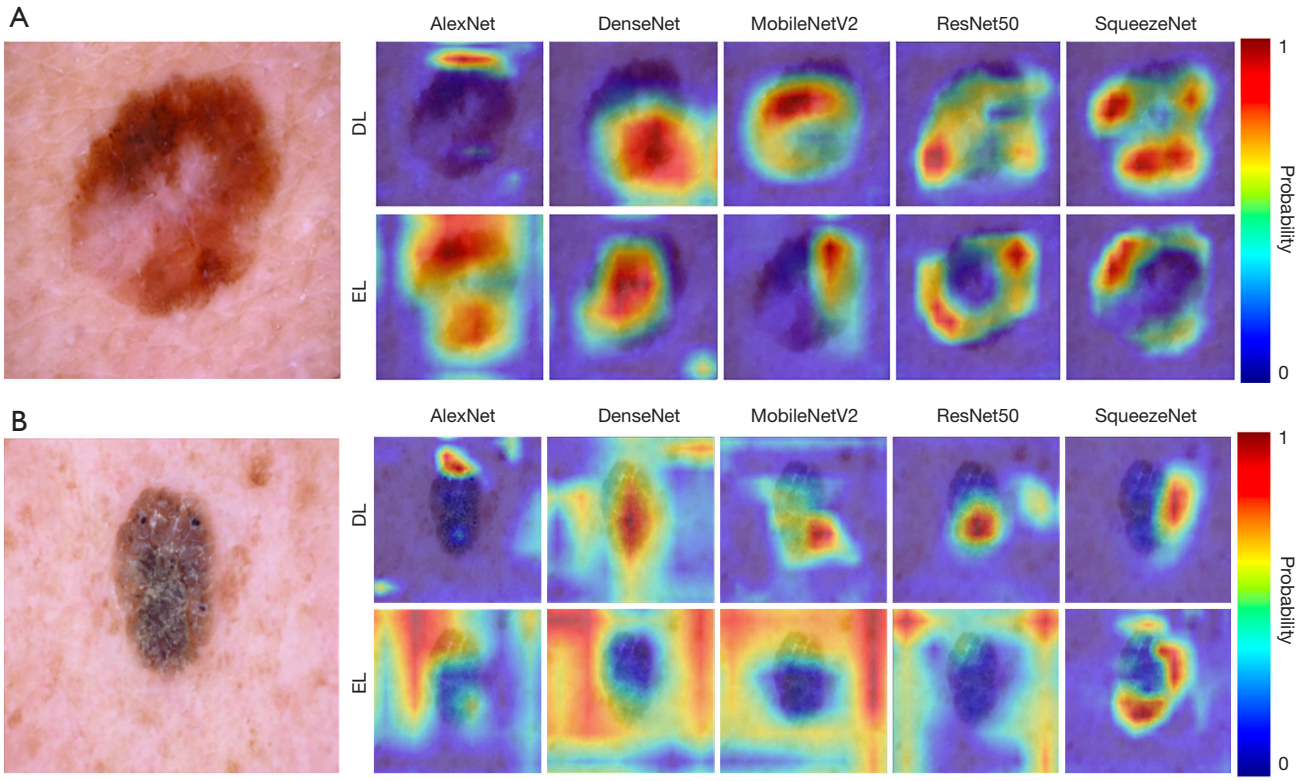

Figure 8 Grad-CAM outputs outlining regions of differential features contributing towards network classification of malignant features. EL networks more accurately isolate key regions of nevi that serve as markers of malignancy. (A) Grad-CAM images of N_i and E_i on melanoma lesions. (B) Grad-CAM images of N_i and E_i on benign keratosis-like lesions. DL, deep learning; EL, ensemble learning; Grad-CAM, gradient-weighted class activation mapping; E_i, all deep ensemble networks; N_i, all deep networks.

Table 7 One-sided Wilcoxon test demonstrates a significant increase in E_i predictions compared to N_i predictions on $D_{holdout}$

Model	P value
AlexNet	2.07E−19
DenseNet	1.27E−05
MobileNetV2	7.91E−07
ResNet50	1.49E−75
SqueezeNet	7.77E−04

E_i, all deep ensemble networks; N_i, all deep networks; $D_{holdout}$, Holdout Test Dataset.

as an additional feature involved in decision-making, E_i makes accurate decisions by highlighting specific features of the nevus. On the other hand, N_i yielded notable classification performance, but the Grad-CAM images demonstrate less specificity of malignant features to inform a decision. More importantly, however, ensemble networks in E_i demonstrate the capability of informing a holistic decision by taking into account the features of the entire image.

A one-sided Wilcoxon test (48) was employed to assess significant differences between network predictions on $D_{holdout}$ for N_i and E_i. Paired prediction data for N_i and E_i was compared to generate two-tailed differences in prediction. Across all network architectures, we noticed a significant increase in the diagnostic accuracy of predictions on $D_{holdout}$ by E_i (*Table 7*).

We tracked network agreeability across N_i and E_i by evaluating the Pearson's correlation coefficient (R) of pairwise network predictions on $D_{holdout}$. On average, there was a moderate correlation between both modalities, with a pairwise correlation across N_i (R_{DL}) of 0.621 and across E_i (R_{EL}) of 0.658 (*Figure 9*). While there was no significant difference in the average pairwise correlation of predictions (P=0.581), certain pairs of networks demonstrated prominent differences in correlation. For example, the

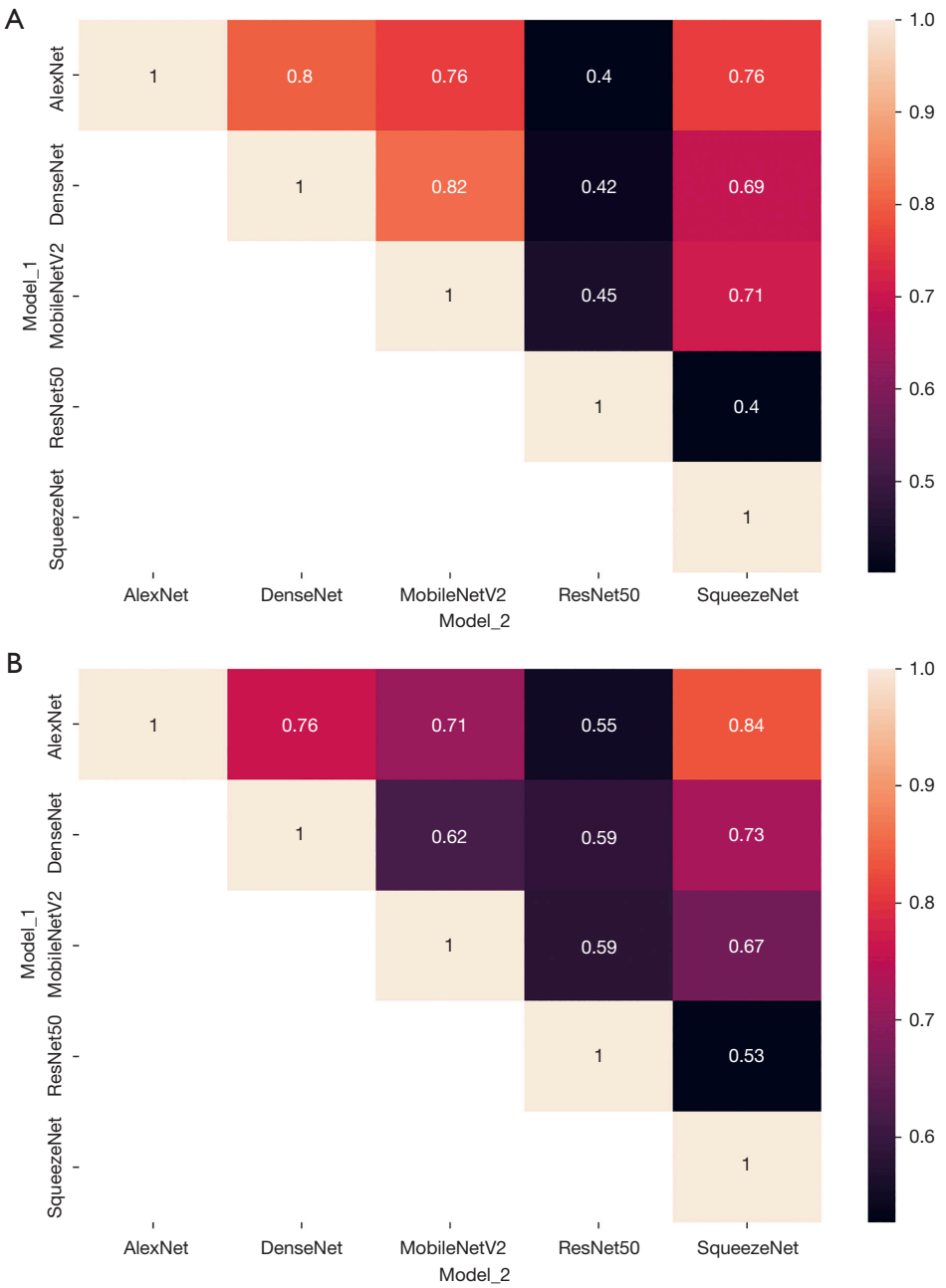

Figure 9 Heatmap of correlation between pairs of network predictions on $D_{holdout}$. (A) Correlation of predictions between models in N_i. (B) Correlation of predictions between models in E_i. $D_{holdout}$, Holdout Test Dataset; N_i, all deep networks; E_i, all deep ensemble networks.

largest difference in correlation between the two modalities was seen across DenseNet and MobileNetV2, where the pairwise correlation of E_D and E_M was 0.2 less than the correlation of N_D and N_M.

Additionally, across both modalities, the ResNet50 model had the poorest pairwise correlation compared to the rest of the network architectures. Upon analyzing predictions, this is due to the higher false positive rate in N_R and E_R. Regardless, the ResNet50 architectures demonstrated no significant difference in AUC compared to the other networks of the same modality, indicating that performance was not impaired despite the stark decrease in

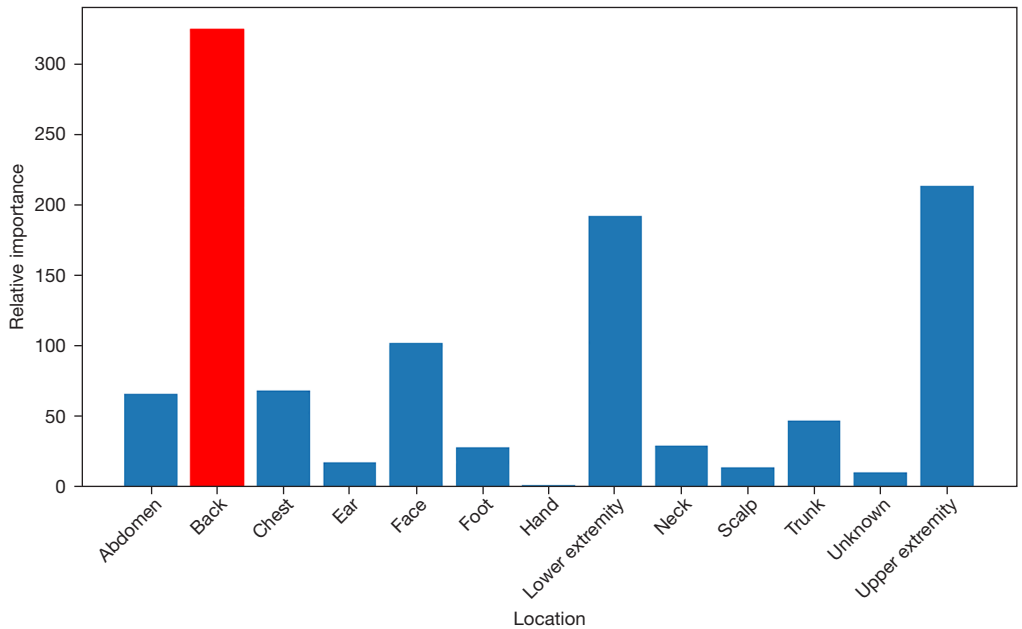

Figure 10 Relative importance of localization in informing M_{RF} classification of melanoma. M_{RF}, Random forest machine learning model.

pairwise correlation.

Discussion

While CNNs have recently demonstrated astronomical increases in the performance of image classification (49,50) and segmentation (51), there is an increasing need to provide parameters to augment their performance. Techniques such as pre-processing (52) and batch normalization (53) have been mainstays in standardizing image quality when passing into CNNs. The relatively new emergence of EL provides a unique modality to increase CNN performance. EL utilizes the outputs of separate networks specific to individual tasks to generate a composite output informed by highly specific networks (54).

Recently, applications of EL networks have shown another dimension of promise in augmenting DL performance for medical imaging diagnostics. Xiao *et al.* developed a series of deep classification algorithms, each specialized to extract differential features of pulmonary nodules within a region of interest and weighted predictions to determine the malignancy of chest computed tomography scans. Their proprietary integrated classifier yielded a testing accuracy of 93.10%±2.4% when using this ensemble deep network (55). The work of Dmitriev *et al.* proved to be very similar to our approach, where they combined the output of a Bayesian random forest algorithm trained on demographic data with a novel CNN to classify the four most common pancreatic cysts from computed tomography images. Their novel approach underwent 10-fold CV to yield a testing accuracy of 83.6% (56).

The application of EL for dermatoscopic skin lesions proved to have a significantly increased binary classification accuracy when compared to the performance of DL. In the dermatology community, the prevalence of skin cancer cannot only be determined by superficial, phenotypic features; demographic factors play a significant role in this classification challenge. We conducted feature importance extraction for the localization of lesions on M_{RF} (*Figure 10*). From this data, the highest prevalence of melanoma was observed on the back, with lesions on the upper extremity having a similar level of importance extracted from M_{RF}. This result can only be applied to the HAM10000 dataset and more clinical data would be needed to form a conclusion.

This study served as a baseline in demonstrating the improvement of classification performance of melanoma with the inclusion of a ML risk score as an additional feature in deep ensemble training. We utilized historically-stable classification networks to test our hypothesis. EfficientNet (57) is a novel network architecture that has demonstrated impressive capabilities in classification

challenges. Ali *et al.* demonstrated its diagnostic accuracy with EfficientNet base networks and modified derivatives yielding AUCs ranging from 0.96–0.98 (58). Jeyakumar *et al.* (59) tested five modern DL architectures for multi-class classification of the HAM10000 dataset, exemplifying the classification performance of the GoogleNet (60) architecture. The novel architecture yielded an AUC of 0.98, granting near-perfect predictive capabilities. A study conducted by Dr. Orman Salih analyzes the proficiency of reducing the computational burden on any given CNN via implementing the genetic algorithm that eliminates noise generated from other hyperparameters during model hypertuning as well as using a Fully Connected Network-based model, achieving an accuracy of 99% when tested on the HAM10000 dataset (61). Furthermore, lightweight networks, such as the DeepSkinNet by Abhiram *et al.* have demonstrated exceptional multi-class classification performance on the HAM10000 dataset (62). Their novel network yielded a testing classification accuracy of 0.9734 while having significantly fewer parameters than an industry-standard AlexNet. We are motivated to further test the capabilities of integrating demographic data into modern DL architectures in the hopes of augmenting performance through EL.

Due to the HAM10000 dataset including three demographic features for ML risk stratification, performance was relatively poor, achieving maximum test AUCs of 0.710 with the best-performing model. In actuality, clinical data may present more demographic features that can help inform ML performance, leading to more algorithms being developed for skin cancer risk stratification. For example, prior literature suggests ethnicity, quantification of sun exposure, and genetic predispositions (63) play active roles in a patient's incidence of melanoma. Additionally, testing was only performed on the holdout test set taken from the HAM10000 dataset. Being able to test these networks on an independent institution's dermatoscopic data would allow us to quantify the generalizability of our models.

Conclusions

The use of deep EL in medical diagnostics is one of increasing interest. To the best of our knowledge, this remains the first instance of utilizing a ML risk stratifier to inform CNN decisions for the classification of melanoma using dermatoscopic data. By including the risk stratification of a random forest model as an additional feature in the last convolutional layer within CNNs, we observed a statistically significant increase in multiple performance metrics, including AUC and Wilcoxon signed rank test of predictions, for most deep network architectures. A moderate-to-strong pairwise correlation existed across all networks within each modality of training. Most notably, however, the inclusion of a demographic risk stratifier increased the specificity of features used to inform a decision by networks. We noticed fine-tuned isolation of differential features in Grad-CAM images generated by ensemble networks for melanoma lesions and holistic inclusion of healthy skin to deliver a prediction in benign lesions. While the networks used were standard CNNs, this is a significant first step in delivering non-invasive diagnostics for skin lesions using patient demographic data and dermatoscopic lesions. Though the use of AI as the gold standard in medical diagnostics is far from becoming a reality, it has the potential to play breakthrough roles in healthcare. Computer vision applications can provide affordable risk stratification to rural populations or to patient populations whose lifestyle does not allow for regular physician visits. Especially in the field of dermatology, where skin cancer prevalence is multifactorial in nature, mobile solutions to determine the malignancy of certain nevi can serve as a beneficial tool for the vast patient population. We aim to further broaden the results of integrating a ML risk stratifier in DL skin cancer classification by integrating a multi-class classification. Additionally, we intend on collaborating with clinics to obtain more demographic data for more robust ML performance to inform CNN decision-making. Finally, the integration of deep EL to improve state-of-the-art network architectures may further increase classification performance and we intend to test this hypothesis in future studies.

Acknowledgments

This work was supported by the staff at California Dermatology Care. Their insightful comments and suggestions were invaluable in shaping the direction of this work. We are grateful to the participants of the HAM10000 challenge for contributing towards the widely-used public database of dermatoscopic lesions and demographic data. Without this data, our project would not have come to fruition. We would like to acknowledge the contributions of all the participants who generously gave their time and effort to make this study possible.
Funding: None.

Footnote

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-38/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-38/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This original work conformed to the provisions of the Declaration of Helsinki (as revised in 2013). An ethics review board was not necessary for this project due to the nature of public datasets being used and no direct interaction with animal tissue. Patient data was confidentially stripped of identifying features in the public dataset through the institution's proprietary means.

References

1. Wheless L, Black J, Alberg AJ. Nonmelanoma skin cancer and the risk of second primary cancers: a systematic review. Cancer Epidemiol Biomarkers Prev 2010;19:1686-95.
2. Halpern AC, Altman JF. Genetic predisposition to skin cancer. Curr Opin Oncol 1999;11:132-8.
3. Fagundo E, Rodríguez-García C, Rodríguez C, et al. Analysis of phenotypic characteristics and exposure to UV radiation in a group of patients with cutaneous melanoma. Actas Dermosifiliogr 2011;102:599-604.
4. Leonardi-Bee J, Ellison T, Bath-Hextall F. Smoking and the risk of nonmelanoma skin cancer: systematic review and meta-analysis. Arch Dermatol 2012;148:939-46.
5. Buster KJ, You Z, Fouad M, et al. Skin cancer risk perceptions: a comparison across ethnicity, age, education, gender, and income. J Am Acad Dermatol 2012;66:771-9.
6. Binstock M, Hafeez F, Metchnikoff C, et al. Single-nucleotide polymorphisms in pigment genes and nonmelanoma skin cancer predisposition: a systematic review. Br J Dermatol 2014;171:713-21.
7. Segura S, Puig S, Carrera C, et al. Non-invasive management of non-melanoma skin cancer in patients with cancer predisposition genodermatosis: a role for confocal microscopy and photodynamic therapy. J Eur Acad Dermatol Venereol 2011;25:819-27.
8. Kasmi R, Mokrani K. Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule. IET Image Process 2016;10:448-55.
9. Rigel DS, Friedman RJ, Kopf AW, et al. ABCDE--an evolving concept in the early detection of melanoma. Arch Dermatol 2005;141:1032-4.
10. Kim T, Freeny I, Ting W. 28725 Alternating weekly treatment schedule for superficial radiotherapy for nonmelanoma skin cancers. J Am Acad Dermatol 2021;85:AB206.
11. Shriner DL, McCoy DK, Goldberg DJ, et al. Mohs micrographic surgery. J Am Acad Dermatol 1998;39:79-97.
12. Smeets NW, Krekels GA, Ostertag JU, et al. Surgical excision vs Mohs' micrographic surgery for basal-cell carcinoma of the face: randomised controlled trial. Lancet 2004;364:1766-72.
13. Kuryk L, Bertinato L, Staniszewska M, et al. From Conventional Therapies to Immunotherapy: Melanoma Treatment in Review. Cancers (Basel) 2020;12:3057.
14. Diep D, Calame A, Cohen PR. Morphologic Mimickers of Seborrheic Keratoses: Cutaneous Lesions Masquerading as Seborrheic Keratoses. Cureus 2021;13:e18559.
15. Cui W, Fowlis DJ, Bryson S, et al. TGFbeta1 inhibits the formation of benign skin tumors, but enhances progression to invasive spindle carcinomas in transgenic mice. Cell 1996;86:531-42.
16. Pickett H. Shave and punch biopsy for skin lesions. Am Fam Physician 2011;84:995-1002.
17. Muhammad K, Ullah A, Lloret J, et al. Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Trans Intell Transp Syst 2020;22:4316-36.
18. Oosthuizen K, Botha E, Robertson J, et al. Artificial intelligence in retail: The AI-enabled value chain. Australas Mark J 2021;29:264-73.

19. Milana C, Ashta A. Artificial intelligence techniques in finance and financial markets: a survey of the literature. Strateg Change 2021;3:189-209.
20. Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019;113:47-54.
21. Maron RC, Weichenthal M, Utikal JS, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019;119:57-65.
22. Hsu BW, Tseng VS. Hierarchy-aware contrastive learning with late fusion for skin lesion classification. Comput Methods Programs Biomed 2022;216:106666.
23. Zunair H, Ben Hamza A. Melanoma detection using adversarial training and deep transfer learning. Phys Med Biol 2020;65:135005.
24. Li S, Chu Y, Wang Y, et al. Distinguish the Value of the Benign Nevus and Melanomas Using Machine Learning: A Meta-Analysis and Systematic Review. Mediators Inflamm 2022;2022:1734327.
25. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 2020;43:635-40.
26. Coups EJ, Manne SL, Heckman CJ. Multiple skin cancer risk behaviors in the U.S. population. Am J Prev Med 2008;34:87-93.
27. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.
28. Lee T, Ng V, Gallagher R, et al. DullRazor: a software approach to hair removal from images. Comput Biol Med 1997;27:533-43.
29. Freund Y, Mason L. The alternating decision tree learning algorithm. Inicml 1999;99:124-33.
30. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.
31. Murphy KP. Naive bayes classifiers. Univ Br Columbia 2006;18:1-8.
32. Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random forest. InInformation Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3. Heidelberg: Springer Berlin; 2012:246-52.
33. Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. IEEE Intell Syst Their Appl 1998;13:18-28.
34. Böhning D. Multinomial logistic regression algorithm. Ann Inst Stat Math 1992;44:197-200.
35. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 1985;(4):580-5.
36. Kheradpisheh SR, Ghodrati M, Ganjtabesh M, et al. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. Sci Rep 2016;6:32672.
37. Ajmal H, Rehman S, Farooq U, et al. Convolutional neural network based image segmentation: a review. Pattern Recognit Track XXIX 2018;10649:191-203.
38. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput 2017;29:2352-449.
39. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115:211-52.
40. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM 2017;60:84-90.
41. Iandola F, Moskewicz M, Karayev S, et al. Densenet: Implementing efficient convnet descriptor pyramids. ArXiv Prepr ArXiv14041869. 2014.
42. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2018:4510-20.
43. Debgupta R, Chaudhuri BB, Tripathy BK. A wide ResNet-based approach for age and gender estimation in face images. In: International Conference on Innovative Computing and Communications: Proceedings of ICICC, 2019. Singapore: Springer Singapore; 2020;1:517-30.
44. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. IEEE; 2009:248-55.
45. Iandola FN, Han S, Moskewicz MW, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv; 2016 [cited 2023 May 24]. Available online: http://arxiv.org/abs/1602.07360
46. Bisong E. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Berkeley, CA: Apress; 2019 [cited 2023 Jul 18]. Available online: http://link.springer.com/10.1007/978-1-4842-4470-8
47. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2016 [cited 2023 Jul 18]; Available online: https://arxiv.org/abs/1610.02391

48. Cuzick J. A Wilcoxon-type test for trend. Stat Med 1985;4:87-90.
49. Ul Haq A, Li J, Memon MH, et al. Comparative Analysis of the Classification Performance of Machine Learning Classifiers and Deep Neural Network Classifier for Prediction of Parkinson Disease. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). Chengdu, China. IEEE; 2018:101-6. Available online: https://ieeexplore.ieee.org/document/8632613/
50. Roge A, Hiremath A, Sobota M, et al. Evaluating the sensitivity of deep learning to inter-reader variations in lesion delineations on bi-parametric MRI in identifying clinically significant prostate cancer. In: Iftekharuddin KM, Drukker K, Mazurowski MA, et al. editors. Medical Imaging 2022: Computer-Aided Diagnosis. San Diego, United States: SPIE; 2022:41. Available online: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12033/2613245/Evaluating-the-sensitivity-of-deep-learning-to-inter-reader-variations/10.1117/12.2613245.full
51. Işın A, Direkoğlu C, Şah M. Review of MRI-based Brain Tumor Image Segmentation Using Deep Learning Methods. Procedia Comput Sci 2016;102:317-24.
52. Masoudi S, Harmon SA, Mehralivand S, et al. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. J Med Imaging (Bellingham) 2021;8:010901.
53. Thakkar V, Tewary S, Chakraborty C. Batch Normalization in Convolutional Neural Networks — A comparative study with CIFAR-10 data. In: 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT). Kolkata: IEEE; 2018:1-5. Available online: https://ieeexplore.ieee.org/document/8470438/
54. Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33:1-39.
55. Xiao N, Qiang Y, Zia MB, et al. Ensemble classification for predicting the malignancy level of pulmonary nodules on chest computed tomography images. Oncol Lett 2020;20:401-8.
56. Dmitriev K, Kaufman AE, Javed AA, et al. Classification of Pancreatic Cysts in Computed Tomography Images Using a Random Forest and Convolutional Neural Network Ensemble. In: Descoteaux M, Maier-Hein L, Franz A, et al. editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Cham: Springer International Publishing; 2017:150-8. (Lecture Notes in Computer Science; vol. 10435). Available online: https://link.springer.com/10.1007/978-3-319-66179-7_18
57. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv; 2020 [cited 2023 Jul 19]. Available online: http://arxiv.org/abs/1905.11946
58. Ali K, Shaikh ZA, Khan AA, et al. Multiclass skin cancer classification using EfficientNets - a first step towards preventing skin cancer. Neurosci Inform 2022;2:100034.
59. Jeyakumar JP, Jude A, Priya Henry AG, et al. Comparative Analysis of Melanoma Classification Using Deep Learning Techniques on Dermoscopy Images. Electronics 2022;11:2918.
60. Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions. 2014 [cited 2023 Jul 19]; Available online: https://arxiv.org/abs/1409.4842
61. Salih O, Duffy KJ. Optimization Convolutional Neural Network for Automatic Skin Lesion Diagnosis Using a Genetic Algorithm. Appl Sci 2023;13:3248.
62. Abhiram AP, Anzar SM, Panthakkan A. DeepSkinNet: A Deep Learning Model for Skin Cancer Detection. In: 2022 5th International Conference on Signal Processing and Information Security (ICSPIS). Dubai, United Arab Emirates: IEEE; 2022:97-102. Available online: https://ieeexplore.ieee.org/document/10002541/
63. Matthews NH, Li WQ, Qureshi AA, et al. Epidemiology of Melanoma. In: Ward WH, Farma JM. editors. Cutaneous Melanoma: Etiology and Therapy. Codon Publications; 2017:3-22. Available online: https://exonpublications.com/index.php/exon/article/view/168

doi: 10.21037/jmai-23-38
Cite this article as: Roge A, Ting P, Chern A, Ting W. Deep ensemble learning using a demographic machine learning risk stratifier for binary classification of skin lesions using dermatoscopic images. J Med Artif Intell 2023;6:14.