Image Quality Assessment Using Convolutional Neural Network in Clinical Skin Images

Hyeon Ki Jeong[1], Christine Park[2], Simon W. Jiang[2], Matilda Nicholas[3], Suephy Chen[3,4], Ricardo Henao[1] and Meenal Kheterpal[3]

The image quality received for clinical evaluation is often suboptimal. The goal is to develop an image quality analysis tool to assess patient- and primary care physician–derived images using deep learning model. Dataset included patient- and primary care physician–derived images from August 21, 2018 to June 30, 2022 with 4 unique quality labels. VGG16 model was fine tuned with input data, and optimal threshold was determined by Youden's index. Ordinal labels were transformed to binary labels using a majority vote because model distinguishes between 2 categories (good vs bad). At a threshold of 0.587, area under the curve for the test set was 0.885 (95% confidence interval = 0.838–0.933); sensitivity, specificity, positive predictive value, and negative predictive value were 0.829, 0.784, 0.906, and 0.645, respectively. Independent validation of 300 additional images (from patients and primary care physicians) demonstrated area under the curve of 0.864 (95% confidence interval = 0.818–0.909) and area under the curve of 0.902 (95% confidence interval = 0.85–0.95), respectively. The sensitivity, specificity, positive predictive value, and negative predictive value for the 300 images were 0.827, 0.800, 0.959, and 0.450, respectively. We demonstrate a practical approach improving the image quality for clinical workflow. Although users may have to capture additional images, this is offset by the improved workload and efficiency for clinical teams.

Keywords: Computer-aided diagnosis, Deep learning, Image Quality Assessment, Medical imaging, Teledermatology

JID Innovations (2024);**4:**100285 doi:10.1016/j.xjidi.2024.100285

INTRODUCTION

Since the onset of the pandemic, the use of teledermatology for patients has increased (Yeboah et al, 2021). These consultations are typically conducted through mobile applications that require patients to capture images of their skin lesions using mobile devices such as smartphones or tablets, which are then sent to dermatologists for remote diagnosis. However, the quality of the images received is often suboptimal, with up to 50% of patients providing images that are poorly lit, off center, or blurry (Vodrahalli et al, 2021). To better ensure a level of care similar to that of in-person care, high-quality images are essential (Haque et al, 2021; Landow et al, 2014). Recently, healthcare systems are allowing unsolicited patient-derived images to be sent with a clinical concern, leading up to 44% increase in patient messages (Borre and Nicholas, 2022). In addition to the challenges of managing this additional work load, low-quality images are a disproportionate burden because it takes a longer time and more effort to reconcile a plan on the basis of poor images, thus leading to physician burn out and decreased opportunity to generate revenue (Jiang et al, 2023). Hence, there is a need to improve image quality from patient-derived images to improve teledermatology and clinical workflows, improve revenue, and reduce physician burn out.

Two approaches exist for improving low-quality images: image denoising and image quality detection. Image denoising is the process of removing noise or unwanted artifacts from an image to improve visual quality and clarity, whereas image quality detection refers to the assessment of the perceived visual quality of an image based on certain criteria or metrics and determine whether an image is of high or low quality on the basis of human perception. A major limitation of the image denoising method is that it can introduce new artifacts that can obfuscate components of the images that are critical for diagnosis. A major advantage of the image quality detection method is that it can provide real-time feedback when detecting low-quality images directly on the patient's mobile device, in a manner that improves the quality to an acceptable level for clinical decision making. However, it is often challenging to determine a threshold of acceptable image quality because there are several factors that can affect the perceived quality of an image, such as the resolution, color, depth, contrast, and noise. Using human evaluation of image quality as a good standard can also be subjective because each person has a different criterion for good quality image.

[1]*Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA;* [2]*Duke University School of Medicine, Durham, North Carolina, USA;* [3]*Department of Dermatology, Duke University School of Medicine, Durham, North Carolina, USA; and* [4]*Durham VA Medical Center, Durham, North Carolina, USA*

Correspondence: Meenal Kheterpal, Department of Dermatology, Duke University School of Medicine, 40 Duke Medicine Circle, Durham, North Carolina 27710, USA. E-mail: meenal.kheterpal@duke.edu

Abbreviations: AUC, area under the curve; CNN, convolutional neural network; IQA, Image Quality Assessment; NPV, negative predictive value; PPV, positive predictive value

Received 3 July 2023; revised 24 December 2023; accepted 6 March 2024; accepted manuscript published online XXX; corrected published online XXX

Cite this article as: JID Innovations 2024;4:100285

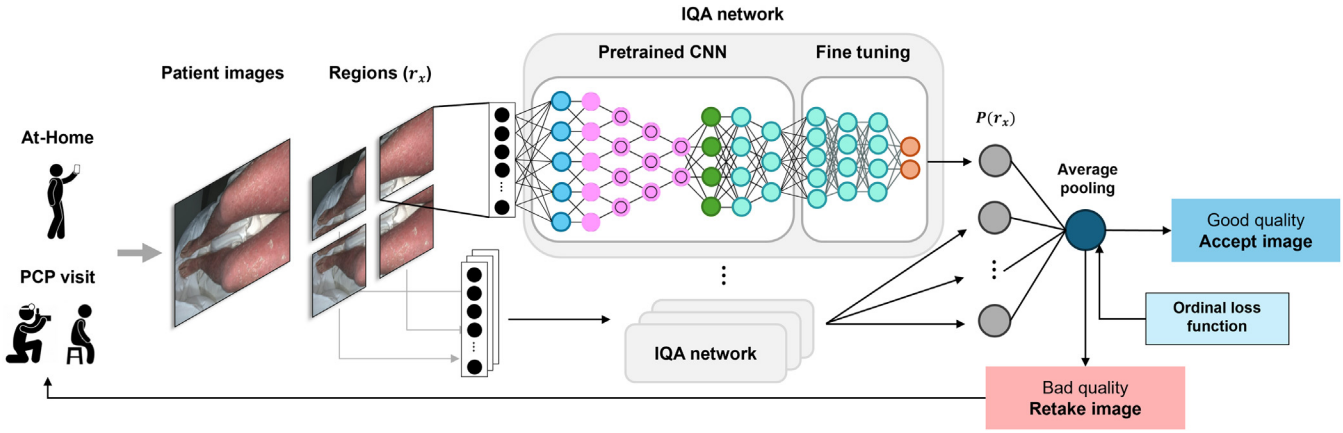

Figure 1. Overview of the IQA network architecture. The input images are partitioned into smaller regions and are processed through the neural network architecture. The resulting outputs are aggregated, and a threshold criterion is applied to determine whether the image is accepted (good quality) or rejected (bad quality). CNN, convolutional neural network; IQA, Image Quality Assessment; PCP, primary care physician.

In this study, we introduce a method using convolutional neural network (CNN) to evaluate image quality of skin photographs. Figure 1 illustrates the overview of the Image Quality Assessment (IQA) process, which includes partitioning images into multiple patches, feeding them into a CNN, and evaluating them against a threshold for the acceptance or rejection of input images.

IQA

Various techniques for evaluating image quality have been proposed. Kim and Lee (2017) introduced DeepIQ, a deep neural network capable of identifying noisy regions within an image and comparing the resulting noise maps with human evaluations (Kim and Lee, 2017). Bianco et al (2018) presented DeepBIQ, a CNN that identifies low-quality images and achieves near-human performance on smartphone photos. Madhusudana et al (2022) developed CONTRIQUE, a contrastive deep learning system for generating transferable representations using unlabeled image quality datasets. However, a common limitation of all these methods is the absence of a reference standard label, which restricts both their training and validation rigor. As a result, unsupervised training approaches are often used, with validation primarily relying on subjective evaluation. In the context of teledermatology, Vodrahalli et al (2021) proposed a classical machine learning image quality classifier. This approach employs automated classical computer vision techniques to identify blur, lighting, and zoom issues in an image and provides patients with explanations for quality assessments. Nonetheless, the method has several shortcomings: (i) it has difficulties with cases in which the background is blurry or has poor lighting; (ii) it cannot detect lesion framing issues; (iii) and it cannot filter out skinless images. Jalaboi et al (2023) also introduced ImageQX, a CNN for IQA with a learning mechanism for identifying the most common poor image quality explanations (eg, bad framing, bad lighting, blur, low resolution, distance issues).

CNN and transfer learning

CNNs are well-suited for image-based tasks owing to their ability to recognize and leverage the spatial and temporal patterns in an image that are useful for making predictions. The advantage of CNNs compared with other traditional machine learning algorithms such as support vector machines, naive bayes, or decision trees is that CNNs are able to automatically extract useful features from raw images without the need for manual feature extraction, and it is robust to small variations in images, making them well-suited for image classification tasks. Building an efficient deep learning model requires both high-quantity and high-quality datasets. However, real-world datasets, especially in medical imaging fields, are complex, making it challenging to process, analyze, and extract meaningful features; may not be well-structured or contain large amount of irrelevant information; and may be time consuming to obtain because they often require expert manual annotation and labeling, which can be an expensive and labor-intensive process. One approach to tackle this problem is through transfer learning, which is a technique in machine learning where a model trained on one task is used as a starting point for a different task. Transfer learning is beneficial because it allows the pretrained model to be fine tuned and reused for any tasks without requiring large amount of labeled data or computational power. Most widely used pretrained models for image-related tasks (eg, Visual Geometry Group [VGG] [Simonyan and Zisserman, 2014[1]], ResNet [He et al, 2016], Inception [Szegedy et al., 2015], etc) have been trained on millions of datasets such as from the ImageNet challenge (Russakovsky et al, 2015). One key consideration when applying transfer learning is the degree of similarity between the source task and the target task and the available data of the target task. Medical images are usually small, and they look different from ImageNet data that are natural color images. Thus, some layers of the pretrained model may be kept frozen, meaning that their weights will not be updated during training. The remaining layers, on the other hand, can be unfrozen, allowing their weights to be adjusted on the basis of the input data of the target task. Morid et al (2021) provides a review of transfer learning in medical image analysis and explores widely used model architecture that was used for each

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556; 2014.

anatomical region. In dermatology, variety of pretrained model has been used for skin lesion classification in early melanoma detection, namely VGGNet (Lopez et al., 2017; Yu et al, 2018), InceptionNet (Cui et al, 2019), ResNet (Hosseinzadeh Kassani and Hosseinzadeh Kassani, 2019), and AlexNet (Hosny et al, 2019).

Multiple instance learning

Advancements in deep learning in recent years have emphasized the need for large amounts of data required to solve complex problems such as image recognition, natural language processing, and speech recognition and more that require high-level understanding and decision making. However, assigning labels or annotating these large datasets is often time consuming and expensive, which may hinder the use of deep learning algorithms in many fields. Multiple instance learning is a type of a weakly supervised learning algorithm where the datasets are arranged in collection of instances (Maron and Lozano-Pérez, 1997), called bags, and the labels are provided for each bag rather than the instances themselves. This method allows to leverage weakly labeled data, which is prominent diverse applications such as medical imaging, document classification, and video/audio processing. Multiple instance learning has been used for IQA in previous studies (Largent et al, 2021; Liang et al, 2021) where the authors have attempted to assess the quality of an image by dividing the image into small instances and aggregating the performance metrics from those regions. This could provide a more comprehensive assessment of the image quality by considering multiple regions of the image instead of relying on a single region. Smaller image regions can capture local image characteristics and variations that might be missed when analyzing the entire image.

Ordinal regression

Ordinal regression is a type of regression analysis that is used to predict an ordinal (ordered)-dependent variable on the basis of 1 or more independent variables. Rank-consistent ordinal regression (Cao et al, 2020) is a variation of ordinal regression that learns to predict the ranking of the dependent variable rather than its exact value. This is achieved by minimizing the rank-consistency loss, which measures the consistency between the predicted ranks of the dependent variable and the true ranks. This ordinal regression approach has a variety of practical applications in medical imaging, especially for quality assessment. Defining quality in binary terms can be challenging because it is inherently subjective. Thus, ordinal label can be applied to mitigate potential biases introduced by individual reviewers.

RESULTS AND DISCUSSION

The distribution of the ordinal quality labels by training, validation, and test set is shown in Table 1. The quality sum represents the number of agreements of being a good quality image.

The optimal threshold was determined on the basis of the training set, and the model's performance was evaluated on the test dataset to ensure that the selected threshold generalizes well to new and unseen data.

The area under the curve (AUC) of the model on the test set is 0.885 with 95% confidence interval of 0.838 and 0.933 and as shown in Figure 2. The performance metrics of the model on test dataset are shown in Table 1 for Youden's index of 0.587.

Further evaluation was performed on the independent validation consisting of 300 images. A total of 150 images taken by physicians are likely to include more high-quality images because the images were taken in a clinical setting by the physicians themselves. The AUC on this dataset is 0.864 (95% confidence interval = 0.818−0.909) as shown in Figure 2, and the metrics based on the Youden's index are shown in Table 2 for all 300 images.

One explanation for a low negative predictive value (NPV) of 0.450 in the independent validation could be the difference in the distribution of positives and negatives in the additional dataset compared with that in the preliminary test set. The ratio of the bad-to-good quality images of the additional 300 dataset is 0.15 to 0.85, whereas the ratio of the training, validation, and test datasets is on average 0.27−0.73 with similar distributions across all 3 datasets. The orange and green line in Figure 3 demonstrates that the model is relatively well-calibrated (slopes 1.27 and 1.15, respectively), whereas the blue line indicates that the predicted probabilities from the model are consistently higher than the actual frequencies of the outcome in the data, suggesting that the model is overconfident on this particular dataset and is not well-calibrated. This implies that an imbalanced dataset with higher number of positive classes could lead to higher false negatives, resulting in lower NPV. Thus, adjusting the distribution of positive and negative cases in an external validation dataset to that of a train, validate, and test dataset is important to ensure that the additional study set is representative of the same population as the training data. We sampled the dataset to match the distribution of positive and negatives cases (Table 2) and showed the improved NPV for the adjusted distribution. This allows the external validation dataset to be representative of the target population.

Table 1. Ordinal Quality Label Distribution

						Independent Set		
Quality Sum	Ordinal Label	Train	Validation	Test	Total	Patient	PCP	Total
0 (none agreed)	[0,0,0,0]	133	19	25	177 (14.8%)	4	14	18 (6%)
1	[1,0,0,0]	107	14	26	147 (12.3%)	19	8	27 (9%)
2	[1,1,0,0]	133	24	26	183 (15.2%)	40	14	54 (18%)
3	[1,1,1,0]	222	18	39	279 (23.2%)	34	23	57 (19%)
4 (all agreed)	[1,1,1,1]	305	45	64	414 (34.5%)	53	91	144 (48%)

Abbreviation: PCP, primary care physician.

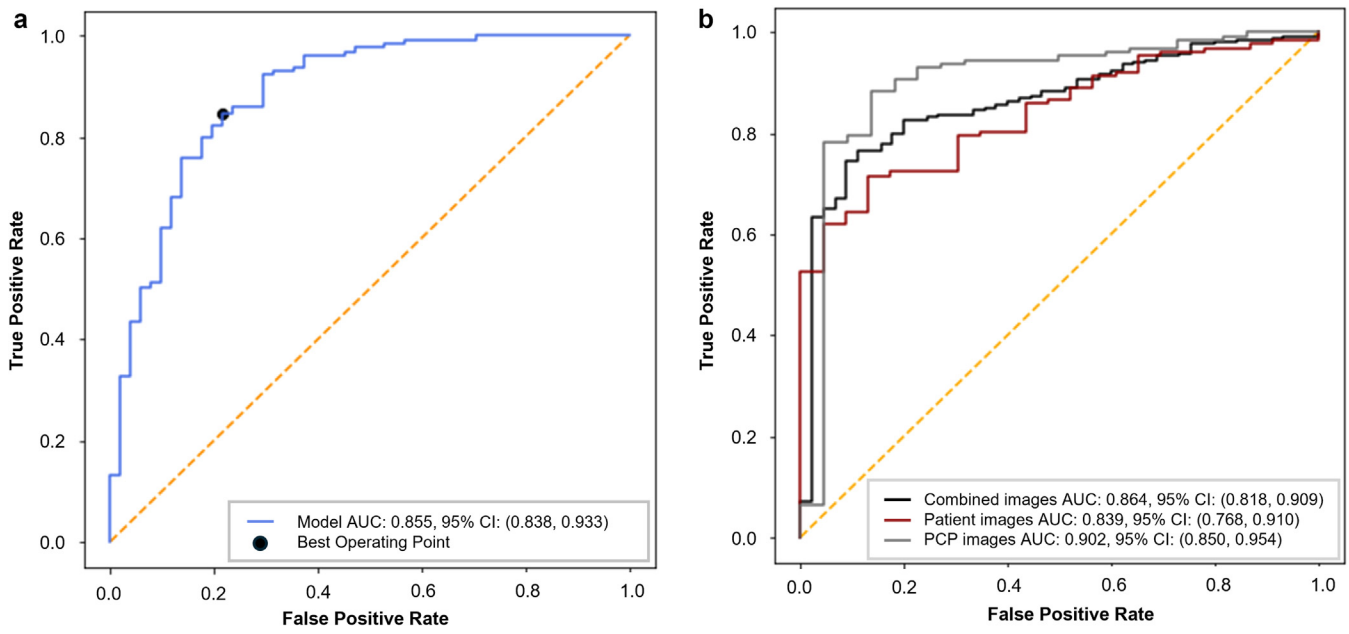

Figure 2. ROC curve of the image quality assessment model on test set and independent set (150 images of patient and PCP images each). (**a**) Test set. **(b)** Independent set. AUC, area under the curve; CI, confidence interval; PCP, primary care physician; ROC, receiver operating characteristic.

The AUC of the model trained on the bounding box was 0.842 with 95% confidence interval of 0.785 and 0.899. The positive predictive value (PPV), true predictive rate, NPV, and true negative rate are 0.851, 0.845, 0.615, and 0.627, respectively. The models using bounding boxed images and the whole images were not statistically significant, suggesting that the inclusion of bounding box definitions did not yield improvements in the model's ability to quantify image quality. Consequently, the additional effort associated with drawing bounding boxes may be deemed unnecessary from a user's perspective.

In this work, we have demonstrated that skin image quality can be assessed using deep learning approach. The model is trained on a dataset of images with ordinal quality labels and evaluated on a separate test dataset and an additional dataset of 300 images. The results show that the model has an AUC of 0.885 on the test set and 0.864 on the independent dataset. The performance metrics of the model such as sensitivity, specificity, PPV, and NPV are reported on the basis of the Youden's index, and the importance of adjusting the distribution of positive and negative cases in external validation datasets is highlighted to ensure that the additional study set is representative of the same population as the training dataset.

Current studies in developing IQA model for skin images include TrueImage by Vodrahalli et al (2021) where they achieved an AUC of 0.759 of defining good versus bad quality image and ImageQX by Jalaboi et al (2023) where they achieved sensitivity and specificity of 0.73 and 0.90. Direct comparison of these works with our model is not possible owing to the different datasets used and different approaches. Our goal was to achieve a PPV of 0.9 to

Table 2. Model's Performance on Test Set

Performance Metrics	Test Set	Additional 300 Images	Adjusted Distribution
AUC	0.885	0.864	0.863 ± 0.022
PPV	0.906	0.959	0.926 ± 0.012
TPR (sensitivity)	0.829	0.827	0.818 ± 0.031
NPV	0.645	0.450	0.627 ± 0.041
TNR (specificity)	0.784	0.800	0.823 ± 0.031

Abbreviations: AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; TNR, true negative rate; TPR, true predictive rate.

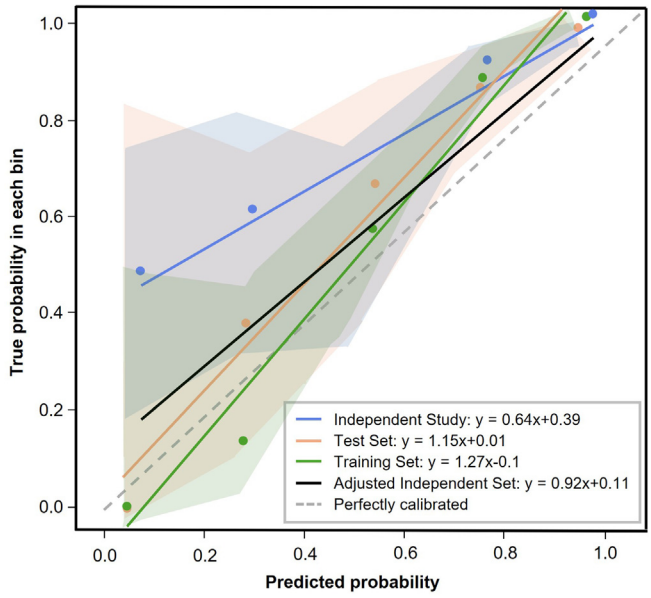

Figure 3. Calibration plot from training set (green), test set (orange), independent study set (blue), and adjusted independent study set (black). IQA, Image Quality Assessment.

optimize clinical workflow. Our model achieved AUC of 0.885 with sensitivity and specificity of 0.829 and 0.784 on the test set and 0.793 and 0.8 on the additional study set, respectively. A lower NPV would predict images that are of good quality (true label) as bad quality (predicted label). This may cause the user to take additional images, despite good quality, which may be useful in the clinical decision making. However, a high PPV will entail a low chance that a bad quality image (true label) will pass as good quality (predicted label), thereby improving clinical workflows and efficiency by reducing workloads through automation of IQA during image acquisition process.

Another important conclusion of our work is the lack of additional useful information for IQA using the region of interest identification, which did not statistically significantly improve the model performance metrics. This entails that good-quality images can be obtained and analyzed with a deep learning approach without additional user effort to highlight the area of interest for a dermatological concern.

Some of the limitations on this work is the number of annotators for the images. To better understand and define the quality of an image, it may take hundreds or up to thousands of reviewers to rate the image either as a binary or continuous scale (Ghadiyaram and Bovik, 2016). Including more experts for annotation can reduce the bias of raters and expertise; however, these works focused mainly on natural images. This may be impractical owing to sensitive nature of medical images provided for dermatological evaluation, although making datasets available for research purposes across institutions can assist with this issue that is faced by medical artificial intelligence. Another limitation is the lack of diverse disease types because we did not have diagnosis information for all the images. We are unable to ensure that images for all disease types will be accurately classified by the model. However, we have images of both inflammatory skin disorders as well as lesions of interest in various anatomical areas, which should capture the majority of diseases. Furthermore, skin color was not considered for the data curation as evidenced by an unbalanced dataset in Table 3. However, the dataset was curated to our local patient population, which serves a specific geographic and demographic area, and the diversity of skin tones in our dataset aligns with the distribution of this population. In this study, the distribution for the training and validation are 76% for Fitz scale 1–3 and 24% for Fitz 4–6. For test set, the distributions are 65% for Fitz 1–3 and 35% for Fitz 4–6. For external dataset, the distributions are 72% for Fitz 1–3 and 28% for Fitz 4–6. The patient demographics for our institute for fiscal year 2023 are in the range of 57–64% for Fitz 1–3 and 28–35% for Fitz 4–6. Although additional images of diverse skin Fitzpatrick types can enhance this model to ensure that it can perform well in all settings, it should be highlighted that the model may still need additional validation studies if population was different within an alternative clinical setting.

Table 3. Fitz Scale

				External Dataset	
Fitz	Train	Validation	Test	Patient	PCP
Type I–III	684	99	117	105	112
Type IV–VI	216	21	63	45	38

Abbreviation: PCP, primary care physician.

Future works will entail real-world validation of the performance of this model within our healthcare system. Additional improvements to further improve the accuracy and reliability of skin image analysis in teledermatology could include additional data with additional annotations, standardizing lighting conditions (especially when tracking disease conditions within the same patient), adding dermoscopy image datasets for quality analysis, and maximizing image resolution metrics. A prospective study to evaluate the utility of the model in our patient population would substantiate real-world use of the model.

MATERIALS AND METHODS

Dataset

The dataset used for this study was acquired from patient images submitted to the Department of Dermatology at Duke University between August 21, 2018 and December 31, 2019, and the details can be found in Jiang et al (2023). This dataset consists of 1200 clinical skin images that were taken by patients using their cell phones or cameras. The images were uploaded to REDCap, and 6 dermatology faculty evaluated 400 images each, 2 dermatology residents evaluated 400 images, and 2 dermatology residents evaluated 200 image quality measures. This assured that each image had 3 different evaluations. One dermatology faculty (MK) evaluated all 1200 images, resulting in 4 total evaluations per image. The dataset images were divided into training, validation, and test sets as 75, 10, and 15% of the dataset, respectively. The images were taken with cell phones in RGB scale, with resolution being 75 pixels per inch on average for both height and width and average pixel size of the image with standard deviation being $2908 \times 2700 \pm 1094 \times 967$ (1360 ± 863 inch^2).

When building a deep learning model in health care, it is often important to have additional datasets beyond training, validation, and test datasets for external validation of the IQA model. In this regard, an external set of 300 images from Duke University Medical Center were acquired, consisting of 150 images taken by the patients themselves and 150 images taken by the primary care physicians, and sent to dermatologists for clinical care decisions. These datasets can assess the model's performance in different contexts and its ability to generalize to new and unseen data. The images were taken with cell phones and iPads in RGB scale, with resolution being 76 pixels per inch on average for both height and width and average pixel size of the image with standard deviation being $2840 \times 2793 \pm 986 \times 983$ (1471 ± 707 inch^2). Thus, once the model has been trained and validated, its performance was evaluated on the test dataset to estimate overall generalization error as well as this external dataset for additional validation. Fitz scale are shown in Table 3 for both training, validation, test, and external dataset.

Model architecture

The proposed IQA model used in this study is the Visual Geometry Group (Simonyan and Zisserman, 2014)[2] as a base model, which have been commonly used for skin analysis in other studies (Morid et al, 2021). After assessing various Visual Geometry Group

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556; 2014.

architectures, VGG16 model, which consists of 13 convolutional layers and 3 fully connected layers, was chosen because it outperformed all other Visual Geometry Group models under consideration. The preprocessing steps for training include dividing the image into 5 regions with equal size (4 corners and the central crop); random vertical and horizontal flip for data augmentation; and resizing each region into 224 × 224 pixels, which is a common size used for many pretrained models. The model was initialized from a VGG16 model pretrained on ImageNet and then fine tuned by freezing the first 5 convolutional layers and unfreezing the remaining 8 as well as the classification head (a single fully connected layer with a sigmoid activation), allowing the model to learn and update part of the weights on the basis of the input data. The model was trained using stochastic gradient descent as the optimizer for 50 epochs with an early stopping on the validation set to prevent overfitting. The ordinal labels were defined as the sum of the reviewer's annotations for each image. The annotation would be 1 if the reviewer agreed that the image was sufficient in quality to be included in the patient chart on medical record where Duke uses EPIC (Verona, WI) and 0 otherwise. These labels were converted into a form of 1-hot encoding where the redefined labels would be a size of N-1 where N represents the total number of classes. Thus, if everyone agreed that the image is of good quality, the ordinal label would be defined as 4, and the 1-hot encoding would be defined as [1,1,1,1], and if everyone disagreed that the image is of bad quality, the ordinal label would be 0, and the 1-hot encoding would be defined as [0,0,0,0] as shown in Table 1.

Performance evaluation

The model's performance was evaluated using AUC of the receiver operating characteristic, which is a diagnostic metric used to evaluate the performance of a binary classification model on distinguishing between positive and negative classes and represents the sensitivity of the model as a function (trade off) of the true positive rate (sensitivity) or false positive rate (1 − specificity). Other metrics are also reported such as PPV, NPV, and true negative rate. The optimal threshold was determined by Youden's index, which represents the best trade off between sensitivity and specificity and balance the number of false positives and false negatives in the classification results (Schisterman et al, 2005). Once the model predicts the rank, the ordinal labels are transformed to binary labels using a majority vote because the goal of the model is to distinguish between 2 distinct categories (good vs bad quality) and not predict quality as a continuous variable. If more than half of the reviewers agree that the image is of good quality, the overall image quality is considered good; otherwise, the quality is considered bad. By applying this transformation, we can evaluate the model's performance on determining whether the model should accept the image as good quality or tell the user to retake the image owing to poor quality.

Additional methodology was tested by defining a bounding box, a box drawn around the region of interest for each image. This methodology would reflect highlighting and isolating particular region of interest that is key for diagnosing and monitoring skin conditions and serves as an attention for IQA model. This would ideally allow the model to learn the features and characteristics related to quality of the region of interest while neglecting other regions that may influence the model's capability to accurately assess quality.

Calibration plot depicts a graphical representation of the relationship between predicted probabilities from a model and the proportion of true probabilities. A perfectly calibrated model would have points that align along a 45-degree line, indicating that the predicted probabilities match the proportion of true outcomes. If a calibration plot shows that the predicted probabilities deviate from the actual outcomes, this may indicate that the model is biased toward 1 class (eg, model may underfit or overfit) or that the dataset is skewed. For example, if the model overpredicts the positive class, this may indicate that the dataset contains a disproportionate number of positive samples.

ETHICS STATEMENT

This study was determined exempt by the Duke University Institutional Review Board (Pro00104856) and included a waiver of informed consent.

DATA AVAILABILITY STATEMENT

Data are sensitive owing to patient images and can be made available upon request (contact Meenal Kheterpal; e-mail: meenal.kheterpal@duke.edu).

ORCIDs

Hyeon Ki Jeong: http://orcid.org/0000-0001-6680-2012
Christine Park: http://orcid.org/0000-0002-0066-366X
Simon Jiang: http://orcid.org/0000-0002-9509-9001
Matilda Nicholas: http://orcid.org/0000-0002-2179-0529
Suephy C. Chen: http://orcid.org/0000-0002-0678-7380
Ricardo Henao: http://orcid.org/0000-0003-4980-845X
Meenal Kheterpal: http://orcid.org/0000-0002-0460-6400

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

The authors acknowledge Jeffery T. Kwock, Krystina Quow, Sarah K. Blanchard, Kimberly F. Breglio, Amber Fresco, Megan O'Brien Jamison, Erin Lesesky, Jane S. Bellet, and Sabrina M. Shearer. This study was funded by Duke AI Health Data Science Microsoft Fellowship Program.

AUTHOR CONTRIBUTIONS

Conceptualization: HKJ, RH, SCC, MK; Funding Acquisition: MK; Project Administration: MK; Supervision: RH, MK; Visualization: HKJ; Writing - Original Draft Preparation: HKJ; Writing - Review and Editing: HKJ, CP, SJ, MN, RH, MK

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE (AI) OR LARGE LANGUAGE MODELS (LLMs)

The author(s) did not use AI/LLM in any part of the research process and/or manuscript preparation.

REFERENCES

Bianco S, Celona L, Napoletano P, Schettini R. On the use of deep learning for blind image quality assessment. SIViP 2018;12:355−62.

Borre ED, Nicholas MW. The disproportionate burden of electronic health record messages with image attachments in dermatology. J Am Acad Dermatol 2022;86:492−4.

Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit Lett 2020;140:325−31.

Cui X, Wei R, Gong L, Qi R, Zhao Z, Chen H, et al. Assessing the effectiveness of artificial intelligence methods for melanoma: a retrospective review. J Am Acad Dermatol 2019;81:1176−80.

Ghadiyaram D, Bovik AC. Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process 2016;25:372−87.

Haque W, Chandy R, Ahmadzada M, Rao B. Teledermatology after COVID-19: key challenges ahead. Dermatol Online J 2021;27. 13030/qt5xr0n44p.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770−8.

Hosseinzadeh Kassani S, Hosseinzadeh Kassani P. A comparative study of deep learning architectures on melanoma detection. Tissue Cell 2019;58:76−83.

Hosny KM, Kassem MA, Foaud MM. Classification of skin lesions using transfer learning and augmentation with Alex-net. PLoS One 2019;14:e0217293.

Jalaboi R, Winther O, Galimzianova A. Explainable image quality assessments in teledermatological photography. Telemed J E Health 2023;29:1342–8.

Jiang SW, Flynn MS, Kwock JT, Liu B, Quow K, Blanchard SK, et al. Quality and perceived usefulness of patient-submitted store-and-forward teledermatology images. JAMA Dermatol 2022;158:1183–6.

Jiang SW, Flynn SM, Borre ED, Nicholas MW. Unsolicited patient images and burnout in dermatology. Clin Exp Dermatol 2023;48:127–8.

Kim J, Lee S. Deep learning of human visual sensitivity in image quality assessment framework. https://openaccess.thecvf.com/content_cvpr_2017/papers/Kim_Deep_Learning_of_CVPR_2017_paper.pdf; 2017. (accessed January 26, 2023).

Landow SM, Mateus A, Korgavkar K, Nightingale D, Weinstock MA. Teledermatology: key factors associated with reducing face-to-face dermatology visits. J Am Acad Dermatol 2014;71:570–6.

Largent A, Kapse K, Barnett SD, De Asis-Cruz J, Whitehead M, Murnick J, et al. Image quality assessment of fetal brain MRI using multi-instance deep learning methods. J Magn Reson Imaging 2021;54:818–29.

Liang D, Gao X, Lu W, Li J. Deep blind image quality assessment based on multiple instance regression. Neurocomputing 2021;431:78–89.

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

Madhusudana PC, Birkbeck N, Wang Y, Adsumilli B, Bovik AC. Image quality assessment using contrastive learning. IEEE Trans Image Process 2022;31:4149–61.

Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Adv Neural Inf Process Syst 1997;10:570–6.

Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 2021;128:104115.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115:211–52.

Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 2005;16:73–81.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA: IEEE; 2015. p. 1–9.

Vodrahalli K, Daneshjou R, Novoa RA, Chiou A, Ko JM, Zou J. TrueImage: a machine learning algorithm to improve the quality of telehealth photos. Pac Symp Biocomput 2021;26:220–31.

Yeboah CB, Harvey N, Krishnan R, Lipoff JB. The impact of COVID-19 on teledermatology: a review. Dermatol Clin 2021;39:599–608.

Yu C, Yang S, Kim W, Jung J, Chung KY, Lee SW, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PLoS One 2018;13:e0193321.