Predicting the Clinical Management of Skin Lesions Using Deep Learning

Kumar Abhishek[1,*], Jeremy Kawahara[1], and Ghassan Hamarneh[1]

[1]School of Computing Science, Simon Fraser University, Canada

{kabhishe, jkawahar, hamarneh}@sfu.ca

[*]Corresponding Author: Kumar Abhishek (kabhishe@sfu.ca)

Abstract

Automated machine learning approaches to skin lesion diagnosis from images are approaching dermatologist-level performance. However, current machine learning approaches that suggest management decisions rely on predicting the underlying skin condition to infer a management decision without considering the variability of management decisions that may exist within a single condition. We present the first work to explore image-based prediction of clinical management decisions directly without explicitly predicting the diagnosis. In particular, we use clinical and dermoscopic images of skin lesions along with patient metadata from the Interactive Atlas of Dermoscopy dataset (1,011 cases; 20 disease labels; 3 management decisions) and demonstrate that predicting management labels directly is more accurate than predicting the diagnosis and then inferring the management decision (11.74% and 9.07% improvement in overall accuracy and AUROC respectively), statistically significant at $p < 0.001$. Directly predicting management decisions also considerably reduces the over-excision rate as compared to management decisions inferred from diagnosis predictions (24.56% fewer cases wrongly predicted to be excised). Furthermore, we show that training a model to also simultaneously predict the seven-point criteria and the diagnosis of skin lesions yields an even higher accuracy (improvements of 5.44% and 0.73% in overall accuracy and AUROC respectively) of management predictions. Finally, we demonstrate our model's generalizability by evaluating on the publicly available MClass-D dataset and show that our model agrees with the clinical management recommendations of 157 dermatologists as much as they agree amongst each other.

Keywords: skin lesion, clinical management, deep learning, classification

1 Introduction

Until a few years ago, the computer-aided diagnosis of skin lesions from images involved extracting the lesion boundary to distinguish it from the surrounding healthy skin (i.e., skin lesion segmentation), followed by calculating features based on rules developed by dermatologists such as the ABCD rule and the CASH rule [1, 2] based on the obtained segmentation, and ultimately using these features to train classical machine learning models (e.g., support vector machines and random decision forests [3–8]) to recommend diagnoses. Since skin lesion segmentation is an intermediate task in the dermatological analysis pipeline, the use of deep learning to predict diagnosis directly from the images, bypassing the segmentation, is now commonplace [9–12] and is evident in other imaging modalities as well [13–17]. We project a similar trend where the model deemphasizes predicting the diagnosis and instead prioritizes producing accurate predictions of the ultimate clinical task (e.g., clinical management).

While deep learning based diagnoses of dermatological conditions from images are reaching the performance levels of medical professionals [9, 10, 18, 19], no work has been published to directly predict the management of the disease. Even in scenarios where the diagnosis is decided by an automated prediction model, the general physician or the dermatologist must still decide on the disease management (be it the treatment plan or some other course of action, e.g., requesting other exams or future follow-ups). Moreover, in some cases, accurately diagnosing the underlying skin condition may not be possible from an image alone. For example, a recent study evaluating the 'majority decision' obtained from over a hundred dermatologists for melanoma classification resulted in a sensitivity of 71.8% with respect to the ground truth diagnosis [20]. Thus, in the case where the visual presentation of a lesion is ambiguous, rather than diagnosing the condition, the correct action may be to perform a biopsy to gain further information. Machine learning-based approaches that classify the underlying skin condition and use the predicted skin condition to directly decide on a disease management (e.g., Han et al. [21]) may not well distinguish among different management decisions that exist within a single class. A management decision (e.g., scheduling a follow-up visit to monitor the skin lesion progression) may even be necessary to confirm a diagnosis (when there is insufficient information within the image), and therefore must precede it. For example, the clinical management decision for a nevi without atypical characteristics may be that no further action is required, whereas for a nevi with atypical characteristics, a dermatologist may opt for a clinical follow up or an excision, which may depend on the severity of the atypical characteristics. Therefore, it is desirable to explore the performance of an artificial intelligence based automatic skin disease management prediction system. Such a system can suggest management decisions to a clinician (i.e., as a second opinion) or directly to patients in under-served communities [22]. Moreover, when there are fewer management decisions to choose from than there are diagnosis classes (since multiple

subsets of disease classes may be prescribed the same course of action), predicting the management decisions is likely a simpler computational problem to address than predicting the diagnosis and then inferring the management.

To the best of our knowledge, we are the first to use deep learning to predict management decisions from skin lesion images without relying on explicit diagnosis predictions. We evaluate our proposed method on the Interactive Atlas of Dermoscopy Dataset [23–25], the largest publicly available database containing both dermoscopy and clinical skin lesion images with the associated management decisions, and show that predicting management decisions directly is more accurate than inferring the management decision from a predicted diagnosis. We also validate our model on the publicly available Melanoma Classification Benchmark (MClass-D) [18, 26] and show that our model exhibits excellent generalization performance when evaluated on data from a different source, and that our model's clinical management predictions are in agreement with those made by 157 dermatologists.

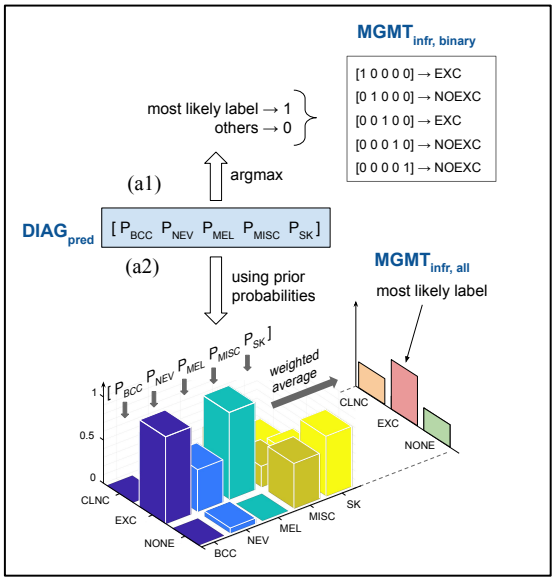
2 Results and Discussion

The Interactive Atlas of Dermoscopy dataset was used to test the performance of a model trained to predict the clinical management decisions ($\text{MGMT}_{\text{pred}}$) compared with inferring the management decisions based on the outputs of a diagnosis prediction model ($\text{MGMT}_{\text{infr}}$). This dataset contains 1,011 lesion cases spanning 20 diagnosis labels (Table 1) grouped into 5 categories [24]: basal cell carcinoma (BCC), nevus (NEV), melanoma (MEL), seborrheic keratosis (SK), and others (MISC), and 3 management decisions: 'clinical follow up' (CLNC), 'excision' (EXC), and 'no further examination' (NONE). The MClass-D dataset [26] was used to compare the diagnosis and the management prediction performance of our model with that of dermatologists. This dataset contains 100 dermoscopic images comprising of 80 benign nevi and 20 melanomas, as well as the responses of 157 dermatologists when asked to make a clinical management decision to each of these 100 images: 'biopsy/further treatment' (EXC) or 'reassure the patient' (NOEXC).

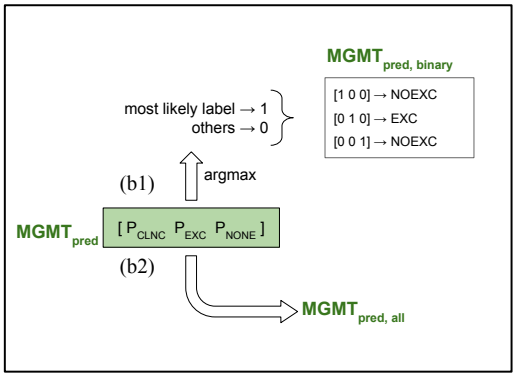
2.1 Interactive Atlas of Dermoscopy Dataset

2.1.1 Predicting whether a lesion should be excised or not

The outputs of the diagnosis prediction model are mapped to a binary management decision ($\text{MGMT}_{\text{infr, binary}}$; Figure 1 (a1)) of whether a lesion should be excised (EXC) or not (NOEXC). All malignancies (MEL and BCC) are mapped to EXC and all other diagnoses to NOEXC. Similarly, the outputs of the management prediction model are mapped to a binary decision ($\text{MGMT}_{\text{pred, binary}}$; Figure 1 (b1)) by retaining the EXC class from $\text{MGMT}_{\text{pred}}$ as is and group-



(a) Predicting the diagnosis to infer the management decision.



(b) Directly predicting the management decision.

(c) Using a multi-task model to simultaneously predict the diagnosis, the management decision, and the 7-point criteria.

Figure 1: An overview of the three prediction models. All the models take the clinical and the dermoscopic images of the skin lesion and the patient metadata as input. (a) The first model predicts the lesion diagnosis probabilities, $\text{DIAG}_{\text{pred}}$. (b) The second model predicts the management decision probabilities, $\text{MGMT}_{\text{pred}}$. (c) The third is a multi-task model and predicts the seven-point criteria ($\text{Criterion}\{1, 2, \cdots, 7\}_{\text{pred, multi}}$) in addition to $\text{DIAG}_{\text{pred, multi}}$ and $\text{MGMT}_{\text{pred, multi}}$. The argmax operation assigns 1 to the most likely label and 0 to all others. For (a), $\text{DIAG}_{\text{pred}}$ diagnosis is used to arrive at a management decision either using (a1) binary labeling, $\text{MGMT}_{\text{infr, binary}}$, or (a2) prior based inference, $\text{MGMT}_{\text{infr, all}}$. Similarly, the outputs of (b) can be used to directly predict a management decision using either (b1) binary labeling, $\text{MGMT}_{\text{pred, binary}}$, or (b2) all the labels, $\text{MGMT}_{\text{pred, all}}$. As explained in the text, the diagnosis labels are basal cell carcinoma (BCC), nevus (NEV), melanoma (MEL), seborrheic keratosis (SK), and others (MISC), and the management decision labels are 'clinical follow up' (CLNC), 'excision' (EXC), and 'no further examination' (NONE). In the case of binary management decisions, we predict whether a lesion should be excised (EXC) or not (NOEXC).

ing CLNC and NONE to form NOEXC. These binary mapping-based approaches serve as our baselines, and we observe that $\text{MGMT}_{\text{infr, binary}}$ correctly predicts 218 of the 395 cases (overall accuracy = 55.19%), whereas $\text{MGMT}_{\text{pred, binary}}$ yields a superior classification performance of 289 correct predictions (overall accuracy = 73.16%), outperforming the inference-based management decision by 17.97%.

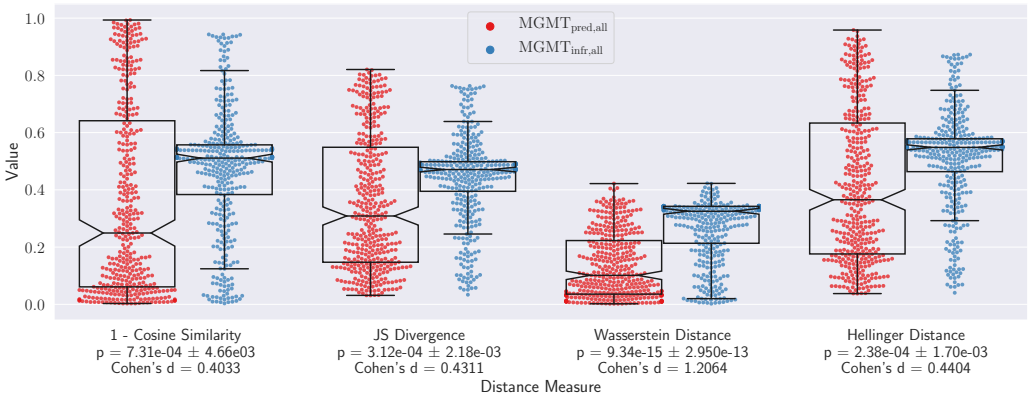
2.1.2 A data-driven approach to inferring management decision from diagnosis predictions

Since cases belonging to a disease label can be managed in multiple ways, a data-driven approach using conditional probabilities (Section 4.3.1 Equation (3)) can be adopted to infer the probabilistic management decisions from the diagnosis predictions, and this does not have to be restricted to a binary management. These inferred management decisions ($\text{MGMT}_{\text{infr, all}}$; Figure 1 (a2)) can then be compared to the probabilistic outputs of the management prediction model ($\text{MGMT}_{\text{pred, all}}$; Figure 1 (b2)).

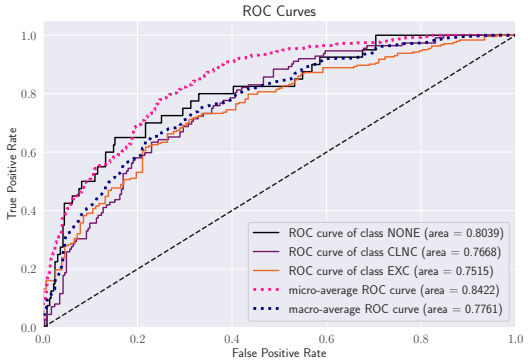
Figure 2 (a) shows the distribution of the four sets of distance measures for examining the correctness of the probabilistic $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ predictions with respect to the target labels, where each dot represents a test case. For (1 - cosine similarity), the mean [95% CI] distance is lower for $\text{MGMT}_{\text{pred, all}}$ as compared to $\text{MGMT}_{\text{infr, all}}$ (0.3584 [0.3260 - 0.3909] versus 0.4703 [0.4490 - 0.4915]; Cohen's d = 0.4033). We observe similar patterns for the Jensen-Shannon divergence (0.3551 [0.3320 - 0.3783] versus 0.4397 [0.4249 - 0.4544]; Cohen's d = 0.4311), the Wasserstein distance (0.1358 [0.1246 - 0.1469] versus 0.2687 [0.2581 - 0.2793]; Cohen's d = 1.2064), and the Hellinger distance (0.4131 [0.3868 - 0.4394] versus 0.5111 [0.4944 - 0.5278]; Cohen's d = 0.4404).

The final management predictions from the two approaches ($\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$) are obtained by extracting the most likely label over the probabilistic predictions, and their quantitative results are presented in Table 2. The ROC curves for the two approaches are shown in Figure 2 (b, c) and their respective confusion matrices, with each cell in the confusion matrices also indicating a diagnosis-wise breakdown of the test samples, are shown in Figure 2 (d, e).

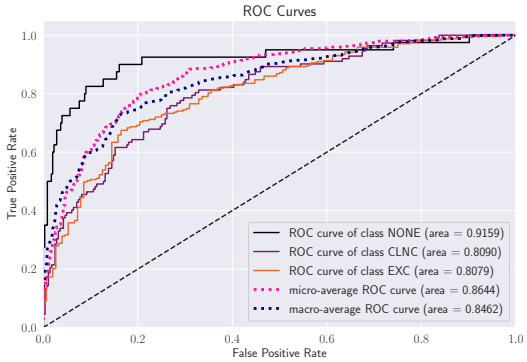
We observe that the overall accuracy and AUROC of $\text{MGMT}_{\text{infr, all}}$ (62.53% and 0.7741) are considerably lower than those of $\text{MGMT}_{\text{pred, all}}$ (69.87% and 0.8443), indicating that predicting the management decisions directly leads to a better accuracy than predicting the diagnosis and then inferring the management. Another interesting observation is that $\text{MGMT}_{\text{infr, all}}$ predictions tend to favor EXC (excision) more than other labels (as can be observed by the dominant blue colored cells in the rightmost column of Figure 2 (b)), which although leads to an excellent sensitivity (0.9835) for the EXC class, yields unacceptable classification performance for the other two classes (0.2 and 0.0, for NONE and CLNC respectively). For example, none of the clinical follow-up cases were predicted correctly by $\text{MGMT}_{\text{infr, all}}$, and 106 cases (94.64%) were predicted to be over-treated by excision. Similarly, the algorithm wrongly predicted excising 32 cases (40%) that, in fact, needed no further examination. On the other hand, $\text{MGMT}_{\text{pred, all}}$ yields a higher overall accuracy without favoring any particular class.



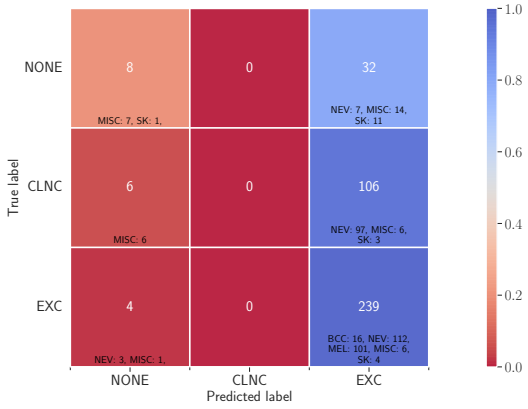
(a) Comparing the distance measures of $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ predictions with the target management labels. Smaller values are better. All differences are statistically significant with $p < 0.001$ (p-values reported below each distance measure).



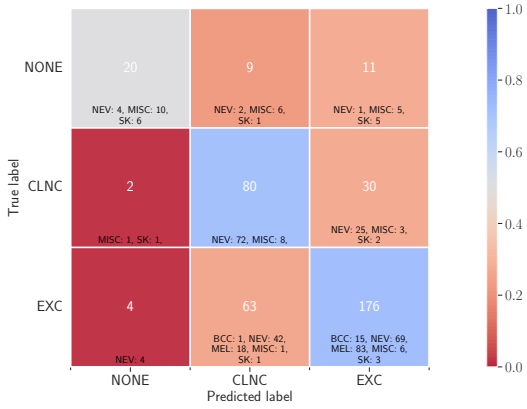
(b) ROC curve of $\text{MGMT}_{\text{infr, all}}$ predictions.



(c) ROC curve of $\text{MGMT}_{\text{pred, all}}$ predictions.



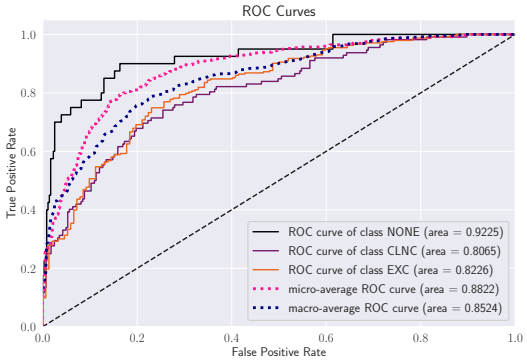
(d) Confusion matrix of $\text{MGMT}_{\text{infr, all}}$ predictions.



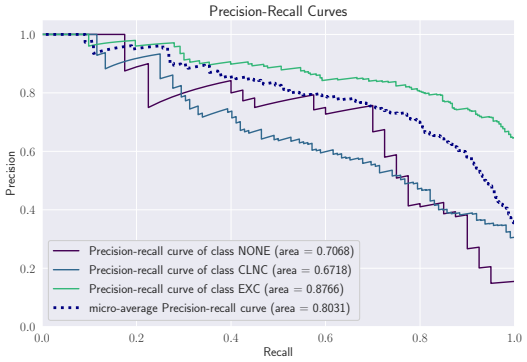
(e) Confusion matrix of $\text{MGMT}_{\text{pred, all}}$ predictions.

Figure 2: Quantitative evaluation of the $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ predictions. (a) Violin plots of the distance measures of the probabilistic predictions show that the $\text{MGMT}_{\text{pred, all}}$ predictions are closer (statistically significant) to the target labels for test data. (b, c) ROC curves and (d, e) confusion matrices of $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ respectively along with cell-wise diagnosis breakdown. Note that $\text{MGMT}_{\text{infr, all}}$ has a tendency to over-excise lesions.

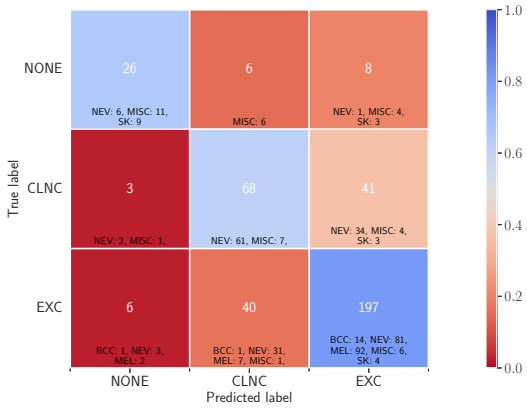
2.1.3 A multi-task prediction model



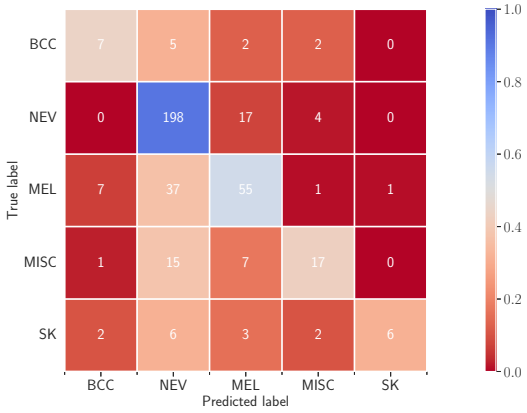
(a) ROC curve of $\text{MGMT}_{\text{pred, multi}}$ predictions.



(b) Precision-recall curve of $\text{MGMT}_{\text{pred, multi}}$ predictions.



(c) Confusion matrix of $\text{MGMT}_{\text{pred, multi}}$ predictions with diagnosis-wise breakdown of the management labels.



(d) Confusion matrix of $\text{DIAG}_{\text{pred, multi}}$ predictions .

Figure 3: Evaluating the multi-modal multi-task model. (a) ROC curve and (b) precision-recall curve for the management prediction task. Confusion matrices for (c) the management prediction task and (d) the diagnosis prediction task along with the diagnosis-wise breakdown for the management labels.

It has been shown that models optimized to jointly predict related tasks perform better than models trained on individual tasks separately [27]. As such, we expect to observe an improvement in the management prediction accuracy of our multi-task model trained to simultaneously predict the seven-point criteria [28] of the lesions ($\text{Criteria1}_{\text{pred, multi}} \cdots \text{Criteria7}_{\text{pred, multi}}$), the diagnosis label ($\text{DIAG}_{\text{pred, multi}}$), and the management decision ($\text{MGMT}_{\text{pred, multi}}$). We plot the confusion matrix and the ROC curves for $\text{MGMT}_{\text{pred, multi}}$ for this multi-task model in Figure 3 (a). As expected, we improve the overall management prediction accuracy by 3.8% (from 69.87% to 73.67%). Moreover, since we have fairly imbalanced classes (see Table 1; for example, there are

243 EXC cases as compared to only 40 NONE cases in the test partition) where ROC curves can indicate an "overly optimistic view" of the algorithm's performance [29], we also plot the precision-recall curves for the multi-task model in Figure 3 (b). A detailed analysis of class-wise performance is presented in Table 3. In addition to its higher management prediction accuracy, this multi-task model may be regarded as less opaque and more trustworthy as its final management prediction was linked to clinically meaningful predictions, i.e., the seven-point criteria and the diagnosis.

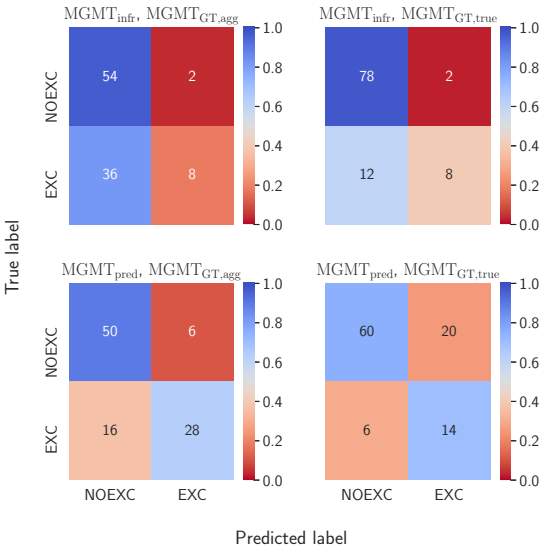
While predicting management decisions, we posit that the clinical penalty of misclassifying certain management decisions is more severe than others. For example, consider a lesion where the correct management decision is for the lesion to be excised. Incorrectly predicting a management decision of 'no further examination' when the lesion should be excised is a more severe mistake than predicting a management decision of 'clinical follow up', since the decision to excise may be corrected in a future examination. We can extend this assumption to also include cases where the model predicts NONE when the target label is CLNC. For example, an EXC or a CLNC misclassified as a NONE is a more severe mistake than a NONE misclassified as an EXC or a CLNC, because in the latter scenario, the best course of action can ultimately be determined by the dermatologist in the clinical visit.

Since the multi-task model has also been trained to predict lesion diagnosis, the confusion matrix for the diagnosis prediction task is shown in Figure 3 (d). Looking at the relationship between the diagnosis and the management labels (Table 1), we notice that all the malignant skin lesions, namely melanomas (MEL) and basal cell carcinomas (BCC), map to the same management label, i.e., excision (EXC). This means that if we can accurately predict a lesion to be either BCC or MEL, we can infer that it has to be excised. Therefore, if we were to first diagnose skin lesions and then infer their management, we would misclassify 46 malignant cases (the number of BCC or MEL misclassified as neither BCC nor MEL; Figure 3 (c)), and thus incorrectly predict their management. On the other hand, if we directly predict the management decisions, we only misclassify 3 malignant cases (1 BCC and 2 MEL; Figure 3 (a)).

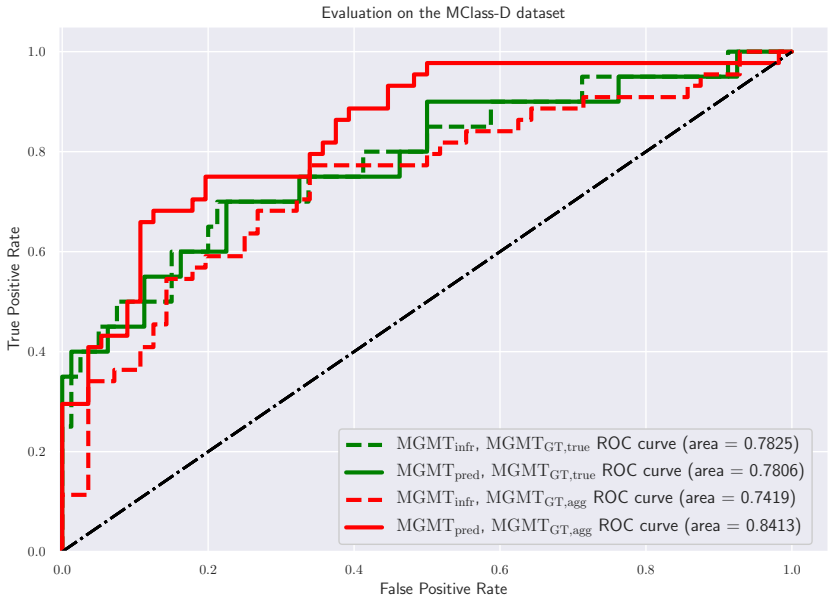
2.2 MClass-D Dataset

Next, we validate our trained prediction model on the publicly available MClass benchmark [26]. For this, we use the multi-task prediction model from Section 2.1.3 to simultaneously predict the diagnosis labels (DIAG) and the clinical management decisions (MGMT) for the 100 dermoscopic images in the MClass-D dataset. We use the multi-task model trained on the Interactive Atlas of Dermoscopy as is and do not fine-tune on the MClass-D dataset.

The prediction classes for DIAG are benign (BNGN) or malignant (MLGN), whereas those



(a) Confusion matrices of $\text{MGMT}_{\text{pred}}$ and $\text{MGMT}_{\text{infr}}$ predictions. The title of each matrix denotes the corresponding predicted and true labels respectively.



(b) ROC curves for $\text{MGMT}_{\text{pred}}$ and $\text{MGMT}_{\text{infr}}$ predictions.

Figure 4: Evaluating the multi-task model on the MClass-D dataset. (a) Confusion matrices and (b) ROC curves for $\text{MGMT}_{\text{pred}}$ and $\text{MGMT}_{\text{infr}}$ predictions with both $\text{MGMT}_{\text{GT, agg}}$ and $\text{MGMT}_{\text{GT, true}}$ as target clinical management labels.

for MGMT are excision (EXC) or not (NOEXC). While the diagnosis ground truth labels from the ISIC Archive are available for the lesions, there are multiple ways of choosing a target label for the clinical management decision. Therefore, we look at two possible ways of assigning the "ground truth" management decision: using the aggregated recommendations of the 157 dermatologists present in the dataset ($\text{MGMT}_{\text{GT, agg}}$), or using the the diagnosis ground truth to derive the "true" management decision ($\text{MGMT}_{\text{GT, true}}$) (where "true" indicates the ideal management decision if the underlying diagnosis was known). For each of the two scenarios, we compare the performance of the directly predicted management decision ($\text{MGMT}_{\text{pred}}$) to that of a scenario when the predicted diagnosis is used to infer the management decision ($\text{MGMT}_{\text{infr}}$), similar to Section 2.1.1.

The confusion matrices and the ROC curves for these two sets of predictions ($\text{MGMT}_{\text{infr}}$ and $\text{MGMT}_{\text{pred}}$) as compared to both methods of choosing the "ground truth" management labels are presented in Figure 4 (a) and (b) respectively. When we set $\text{MGMT}_{\text{GT, agg}}$ as the target labels, as shown in the left column and red curves of Figure 4 (a) and (b) respectively, we observe that predicting the management decision directly ($\text{MGMT}_{\text{pred}}$) performs well for both the management classes without favoring any single particular class and achieves a notable improvement in the area under the ROC curve, as compared to when inferring the management decision ($\text{MGMT}_{\text{infr}}$) based on the model's diagnosis prediction. Additionally, as discussed in Section 2.1.3 , not all misclassification errors are equal, and the clinical penalty of misclassifying an EXC as NOEXC is much more than other errors. While an ROC curve shows the performance over all probability thresholds, the AUROC does not consider the actual decision of the model. When using a default probability threshold of 0.5, we note that directly predicting the management decisions incurs far fewer such mistakes than inferring the management (16 versus 36). Similarly, when setting $\text{MGMT}_{\text{GT, true}}$ as the target management labels, we observe that although the area under the ROC curves are similar (Fig. 4 (b) green curves), the confusion matrix (Fig. 4 (a) right column) reveals that the $\text{MGMT}_{\text{pred}}$ leads to better overall performance across both the classes and fewer instances of EXC being misclassified as NOEXC (6 versus 12).

For evaluating the agreement between the model's predictions and those of the 157 dermatologists, we calculate two agreement measures - Cohen's kappa and Fleiss' kappa. The Cohen's kappa between our model's predictions and that of the aggregated recommendations of the 157 dermatologists is 0.5424. This is higher than that of the agreement between all pairs of dermatologists (0.4124 ± 0.1032), and is comparable to the agreement between one dermatologist and the aggregated recommendations of all the others, repeated for all dermatologists (0.5497 ± 0.0899). Next, the Fleiss' kappa for agreement among the recommendations of 157 dermatologists is 0.4086. To calculate the Fleiss' kappa for capturing the agreement between our model's predictions with those of the dermatologists, we calculate the agreement among a

set of 156 dermatologists' recommendations and the model's predictions, and repeated by leaving out one dermatologist at a time, yielding a score of 0.4080 ± 0.0006. To address concerns that the recommendations of 156 dermatologists might overshadow the model's predictions in the score calculated above, we repeat this experiment for a set of 10 predictions, comprising of 9 dermatologists' recommendations and the model's predictions, and repeat this 1000 times, yielding a score of 0.3961 ± 0.0301. These results indicate that our model's clinical management predictions agree with those made by dermatologists as much as they do amongst each other.

Although Brinker et al. [18] achieve a better performance at classifying melanomas than our model, we believe this can be attributed to multiple factors. First, Brinker et al. trained their prediction model on over 12,000 images and reported the mean of the results obtained from 10 trained models. Our model, on the other hand, is trained on considerably fewer images (413) and the reported results are from a single training run. Second, the training, validation, and testing partitions for Brinker et al. all come from the same data source, i.e., the ISIC Archive, whereas our model was trained on the Interactive Atlas of Dermoscopy and evaluated on images from the ISIC Archive, leading to a domain shift. CNNs have been shown to exhibit poor generalizability for skin lesion classification tasks when trained and evaluated on separate datasets [30]. Despite this, our multi-task prediction model is able to adapt to the new domain and exhibits strong generalization performance for clinical management predictions.

Limitations

Although this study provides a proof of concept of the potential advantages of using deep learning to directly predict the clinical management decisions of skin lesions over inferring management decisions based on predicted diagnosis labels, it suffers from some limitations. First, the dataset that our model is trained on, the Interactive Atlas of Dermoscopy, only contains 20 diagnosis labels and 3 management labels and is not an exhaustive list of all diagnosis and management decisions. Second, although we trained the models on the Interactive Atlas of Dermoscopy with a reasonable effort on hyperparameter tuning and fine tuning, we did not pursue maximizing the classification accuracy. This means that even though our trained prediction model performs well on a held-out test set and is also able to generalize well when evaluated on data coming from a different source than the one it was trained on, better classification performance may be achievable with careful optimization of the prediction models. Finally, we acknowledge that unlike a dermatologist who has access to richer and non-image patient metadata such as patient history, demographics, patient preferences, and difficulty of diagnosis, our model only makes predictions based on the attributes present in these two datasets. However, this is not a technical limitation of our approach and rich multi-modal patient information can be incorporated as and when such attributes become available.

3 Conclusion

In this work, we proposed a model to predict the management of skin lesions using clinical and dermoscopic lesion images and patient metadata. We showed that predicting the management decisions directly is significantly more accurate than predicting the diagnoses first and then inferring the management decision. Moreover, we also observed a considerable increase in the management prediction accuracy with a multi-task model trained to simultaneously predict the seven-point criteria, the diagnoses, and the corresponding management labels. Furthermore, evaluation of our model on another dataset showed excellent cross dataset generalizability and strong agreement with the recommendations of dermatologists.

Our goal with this work is not to propose a method that overrides the dermatologists, rather to provide a second opinion. Deep learning-based approaches for diagnosis, although commonplace as a clinical tool now [31–33], were far from it a decade ago, and we predict a similar shift towards automated algorithms recommending the clinical management of diseases. Since we have proposed a learning-based approach, the model's predictions can be made more robust and similar to dermatologists' predictions by leveraging more complex patient attributes. Future research directions would include collecting and testing on other datasets with other skin conditions and treatments to assess the value of directly predicting management labels and deemphasizing the latent tasks such as diagnosis prediction.

4 Materials and Methods

4.1 Dataset

We have adopted the Interactive Atlas of Dermoscopy dataset [24], a credible and extensively validated dataset that has been widely used to teach dermatology residents [34–36], to train and evaluate our prediction models. The dataset contains clinical and dermoscopic images of skin lesions, patient metadata (patient gender and the location and the elevation of the lesion), the corresponding seven-point criteria [28] for the dermoscopic images, and the diagnosis and the management labels for 1011 cases with mean [standard deviation] age of 28.08 [18.70] years; 489 males (48.37%); 294 malignant cases (29.08%); skin lesion diameter of 8.84 [5.39] mm. We split the dataset into training, validation, and testing partitions in the ratio of approximately $2:1:2$ ($413:203:395$ to be precise) and maintain a similar distribution of the management labels across all the three subsets. A breakdown of the dataset according to the management and the diagnosis labels along with the details of the three splits is presented in Table 1, and more detailed breakdowns of the dataset according to the diagnosis classes and the patient metadata is presented as Supplementary Information (Supplementary Tables 1 and 2 respectively). We

also present the evaluation of the multi-task prediction model on the MClass-D dataset [18], a collection of 100 dermoscopic images from the ISIC Archive with the corresponding diagnosis labels and the clinical management decision of 157 dermatologists surveyed. The dermatologists came from 12 university hospitals in Germany and 43.9% of them were board-certified. The melanomas in the dataset were histopathology-verified and the nevi were diagnosed as benign either by expert consensus or by a biopsy.

4.2 The prediction models

In this section, we present three management prediction models, a detailed breakdown of which is presented in Figure 5. In order to train prediction models that leverage both the clinical and the dermoscopic images as well as the patient metadata available in the dataset, we use a multi-modal framework [24] and train two models: the first to predict the diagnosis and the second to predict the management decision. For both of these models, we adopt an InceptionV3 [37]-backbone pretrained on the ImageNet dataset [38] as the feature extraction model and drop the final output layer. We combine the extracted features from both clinical and dermoscopic images and compute the global average pooled responses, to which we then concatenate the patient metadata as a one-hot encoded vector. Next, we add a convolutional layer for the prediction task (either the diagnosis or the management), and a final classification layer with the associated loss. We use the categorical cross-entropy loss to train the model, and they are denoted by L_{DIAG} and L_{MGMT} for the diagnosis and the management prediction models respectively. Since there is an inherent class imbalance in the dataset, we adopt a mini-batch sampling and weighting approach [24]. The loss function used to train these two single prediction task models is as follows:

$$L_{<\text{task}>} \equiv L\left((x_c, x_d, x_m), y_{<\text{task}>}|\Theta\right) = -\frac{1}{|b|}\sum_{i=1}^{|b|}\sum_{j=1}^{n_{<\text{task}>}} w_j \cdot y^{(i)}_{<\text{task}>,j} . \log\left(\phi\left(x^{(i)}|\Theta\right)_j\right), \quad (1)$$

where x_c, x_d, x_m denote the clinical image, the dermoscopic image, and the patient metadata, respectively, $|b|$ denotes the size of the mini-batch, 'task' denotes either the diagnosis or the management prediction task, and $y_{<\text{task}>}$ and $n_{<\text{task}>}$ denote the target variable and the number of classes for the corresponding tasks respectively. w_j denotes the weight assigned to the j^{th} class (calculated similar to Kawahara et al. [24]), and $\phi\left(x^{(i)}|\Theta\right)_j$ denotes the predicted probability for the j^{th} class given an input $x^{(i)}$ by the model parameterized by Θ.

It has been shown that models optimized to jointly predict related tasks perform better on the individual tasks than models trained on each individual tasks separately [27, 39]. Therefore, we train a third model by extending the multi-modal multi-task framework [24] to simultaneously

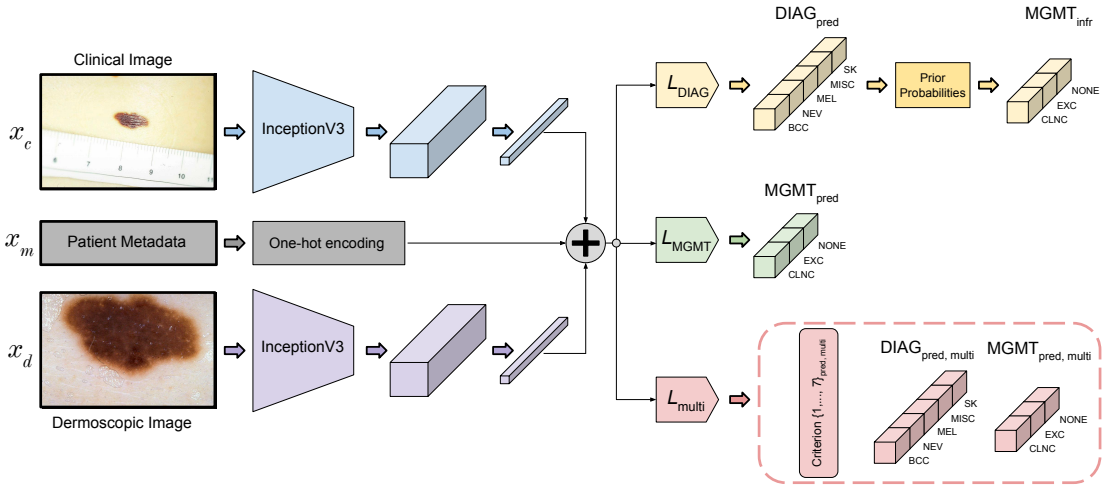


Figure 5: A breakdown of the inputs, outputs, loss functions, and architecture of the three prediction models. Global average pooled feature responses from the clinical and the dermoscopic images are extracted and concatenated (denoted by the plus symbol) with one-hot encoded patient meta-data, and the three models are trained with L_{DIAG}, L_{MGMT}, and L_{multi} respectively. The first model predicts the diagnosis labels ($\text{DIAG}_{\text{pred}}$) which are then used along with the management priors to obtain inferred management decisions ($\text{MGMT}_{\text{infr}}$), whereas the second model predicts the management decisions directly ($\text{MGMT}_{\text{pred}}$). Finally, the last model is a multi-task one and is trained to predict the seven-point criteria, the diagnosis, and the management (outputs enclosed in the dashed box).

predict the seven-point criteria, the diagnosis, and the management decision. The architecture remains the same as the two models described above, except for the last layer, where we add a convolutional layer for each prediction task, and a final classification layer with the multi-task loss. The multi-task loss, denoted by L_{multi} , accounts for all the 9 prediction tasks, namely: the seven-point criteria, the lesion diagnosis, and the lesion management, and is the sum of prediction losses for each of the tasks. As with the previous two models, we adopt the same mini-batch sampling and weighting approach. The loss function used to train this multi-task prediction model is defined as:

$$\begin{aligned} L_{\text{multi}} \equiv L\left((x_c, x_d, x_m), y_{\text{diag}}, y_{\text{mgmt}}, z|\Theta\right) &= L_{\text{DIAG}}\left((x_c, x_d, x_m), y_{\text{diag}}|\Theta\right) \\ &+ L_{\text{MGMT}}\left((x_c, x_d, x_m), y_{\text{mgmt}}|\Theta\right) \\ &+ \sum_{k=1}^{7} L\left((x_c, x_d, x_m), z_k|\Theta\right), \end{aligned} \tag{2}$$

where $L(\cdot)$ denotes the categorical cross entropy loss (as described in Equation (1)) and Θ denotes the parameters of the multi-task model. The model outputs are y_{DIAG}, y_{MGMT}, and

$z_k \in \mathbb{Z}^7$, which denote, respectively, the diagnosis label, the management label, and the integer score for each of the seven-point criteria.

4.3 Making management predictions

4.3.1 Interactive Atlas of Dermoscopy Dataset

Since we ultimately seek the management decision for each patient, we evaluate all the models based on their management prediction performance. We examine two types of management decisions: predicting whether a lesion should be excised or not (our baseline) and predicting all the management decisions. The first model (Figure 1 (a)) is trained to predict the diagnosis, and so we infer the management decisions $\text{MGMT}_{\text{infr}}$ from its diagnosis predictions ($\text{DIAG}_{\text{pred}}$) either by predicting the binary management decision $\text{MGMT}_{\text{infr, binary}}$: EXC versus NOEXC (Figure 1 (a1)), or by predicting all management decisions $\text{MGMT}_{\text{infr, all}}$, which for our dataset are EXC, CLNC, and NONE (Figure 1 (a2)). The second model is trained to predict the management decisions $\text{MGMT}_{\text{pred}}$, either binary $\text{MGMT}_{\text{pred, binary}}$ (Figure 1 (b1)) or all decisions $\text{MGMT}_{\text{infr, all}}$, directly (Figure 1 (b2)). As for the third model, since it is trained to predict the diagnosis and the management along with the 7-point criteria (Figure 1 (c)), we follow the same approach as the first two models to obtain management predictions.

The binary management decisions, $\text{MGMT}_{\text{infr, binary}}$ (Figure 1 (a1)) and $\text{MGMT}_{\text{pred, binary}}$ (Figure 1 (b1)), are obtained using a binary mapping as described in Results and Discussion. Next, given that there are multiple ways to manage a disease category (e.g., in Table 1, NEV cases are managed using all three management labels), we adopt a data-driven approach (Figure 1 (a2)) to calculate the likelihood of all management decisions given a diagnosis prediction. We use the distribution of the management decisions across diagnosis classes in the training data to estimate the priors for assigning a management class m_i to a patient assigned the diagnosis class d_j. This can be denoted as $p(\text{MGMT} = m_i|\text{DIAG} = d_j)$. At inference time, given a patient's data x, we estimate the probability of management m_i by marginalizing over all possible diagnosis classes:

$$P(\text{MGMT} = m_i|x) = \sum_{d_j} p(\text{DIAG} = d_j|x) \; . \; \underbrace{p(\text{MGMT} = m_i|\text{DIAG} = d_j)}_{\text{prior from dataset}} . \tag{3}$$

4.3.2 MClass-D Dataset

The multi-task model used to evaluate the images from MClass-D predicts both the lesion diagnosis ($\text{DIAG}_{\text{pred}}$) and the clinical management ($\text{MGMT}_{\text{pred}}$). The management labels inferred ($\text{MGMT}_{\text{infr}}$) from the diagnosis predictions are obtained by the binary mapping described in Section 2.1.1. To recap, a lesion predicted to be malignant (MLGN) is mapped to the 'excise'

(EXC) label and a lesion predicted to be benign (BNGN) would be mapped to 'do not excise' (NOEXC), meaning that the inferred management decision ($\text{MGMT}_{\text{infr}}$) would have a direct mapping from the predicted diagnosis ($\text{DIAG}_{\text{pred}}$).

Next, we look at the two different ways of obtaining the "ground truth" management labels. First, we aggregate the recommendations of the 157 dermatologists by majority voting to obtain a single prediction for each image ($\text{MGMT}_{\text{GT, agg}}$), and use these as one type of target labels to compare the directly predicted management decisions ($\text{MGMT}_{\text{pred}}$) and the inferred management decisions ($\text{MGMT}_{\text{infr}}$). The second type of target labels are formed by generating the "true" clinical management labels by using a direct mapping from the disease diagnosis to clinical management. This is supported by the fact in an ideal world, we would want all malignancies (MLGN) to be excised (EXC) and all the benign lesions (BNGN) to not be (NOEXC). As such, the "true" clinical management labels ($\text{MGMT}_{\text{GT, true}}$) are obtained by directly mapping the ground truth diagnosis classes to the corresponding management labels.

4.4 Evaluation

For comparing the performance of the baseline binary labeling approach, we compare the per-class sensitivity averaged over the two classes for the two sets of binary predictions, $\text{MGMT}_{\text{infr, binary}}$ and $\text{MGMT}_{\text{pred, binary}}$.

Next, for each of the two sets of management predictions ($\text{MGMT}_{\text{pred, all}}$ and $\text{MGMT}_{\text{infr, all}}$), we obtain probabilistic predictions. To compare the performance of the two models, we choose to evaluate using two methods: (a) using the probabilistic management predictions, and (b) using the most likely label (i.e., choosing the single label with the highest predicted probability). While the evaluation for the latter is rather straightforward with accuracy values and confusion matrices, we formulate the following methodology for evaluating the quality of the probabilistic management predictions: given a set of predicted probability values (over management classes) and the corresponding target management labels, we report distance measures between the probabilistic predictions and the one-hot encoded representations of the target management labels.

4.5 Statistical analysis

The primary outcome measures are class-wise sensitivity, specificity, precision, AUROC and overall accuracy for the diagnosis and the management prediction tasks.

To compare the probabilistic predictions for the management decision obtained using $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$, we use four distance measures to compare the similarity of these probability-

vectors to the one-hot-encoded target labels: cosine similarity, Jensen-Shannon divergence, Wasserstein distance, and Hellinger distance. Since a lower value is better for all these metrics except the cosine similarity, we instead use the (1 - cosine similarity) value for consistency across measures and visualize them using a swarm plot overlaid onto a box plot.

We use the two-sided Wilcoxon signed-rank test [40] to compare the two sets of distance measures for each of the four measures since the differences between the two sets cannot be assumed to be normally distributed. We perform bootstrapping [41] and sub-sampling 1000 times [42] with a sample size of N/2 (where N is the size of the test set) with convergence criteria satisfied [43]. For all the distance measures, we report the means and the 95% confidence intervals along with Cohen's d values [44]. Results are considered statistically significant at $p < 0.001$ level.

For evaluation on the MClass-D dataset, we use two inter-rater measures for assessing the similarity of our model's predictions with those of the 157 dermatologists: Cohen's kappa [45] and Fleiss' kappa [46]. For Cohen's kappa, we calculate the agreement between the model's prediction and the labels obtained by aggregating the recommendations of all 157 dermatologists ($\text{MGMT}_{\text{GT, agg}}$), and compare it with the average agreement between any two dermatologists. To account for the variability among the predictions of multiple dermatologists and how this might not be reflected in the aggregated recommendation, we also compare this with the agreement between one dermatologist and the aggregated recommendation of all others, repeating this over all 157 dermatologists in a leave-one-out fashion and report the average agreement. Unlike Cohen's kappa, Fleiss' kappa can assess the agreement among more than two raters, and therefore we first calculate the agreement among all the 157 dermatologists. For calculating the agreement of the model's predictions with those of the dermatologists, we first calculate the Fleiss' kappa for a set of 157 predictions obtained from 156 dermatologists and our model, and repeat this 157 times in a leave-one-out fashion and report the average agreement. However, this could lead to concerns that the agreement among the 156 dermatologists might affect the kappa value, so we further carry out the same experiment but with a set of 10 management decisions obtained from the recommendations of 9 dermatologists sampled at random from the dataset and our model's predictions. We repeat this 1000 times and report the average agreement.

All statistical analyses were performed in Python using NumPy [47], SciPy [48], statsmodels [49], PyCM [50], and scikit-learn [51] libraries, and all visualizations were created in Python using matplotlib [52] and seaborn [53] libraries.

4.6 Implementation Details

The Keras framework [54] was used to implement all the deep learning models. We follow a similar training paradigm as Kawahara et al. [24]. For all the models, the ImageNet-pretrained weights are frozen at the beginning and the models are fine-tuned with a learning rate of 10^{-3} for 50 epochs, followed by iteratively 'un-freezing' one Inception block at a time (starting from the Inception block closest to the output all the way to the second Inception block) and fine-tuning for 25 epochs with a learning rate of 10^{-3}. We use real-time data augmentation using rotations, horizontal and vertical flipping, zooming, and height and width shifts for these initial 275 epochs. Lastly, we turn off data augmentation and fine-tune for 25 epochs. We use stochastic gradient descent with a weight decay of 10^{-6} and a momentum of 0.9 to optimize the weights.

Acknowledgements

K.A. is funded by Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN 06795 through a research assistantship. The authors are thankful to Ben Cardoen of the Medical Image Analysis Lab for discussions on Figure 2 and statistical analyses. The authors are grateful to the NVIDIA Corporation for donating a Titan X GPU used in this research.

Author Contributions

K.A. worked on writing the code, performing the formal analysis and the experiments, and preparing the figures, with support from J.K. and G.H. K.A. worked on writing the initial draft. G.H. worked on supervising the project, with support from J.K. All authors contributed to the design and the evaluation of the algorithm. All authors contributed to writing, reviewing, and editing the manuscript. All authors read and approved the Article.

Competing Interests

G.H. serves as a Scientific Advisor to Triage Technologies Inc., Toronto, Canada, where J.K. and G.H. are minor shareholders ($< 5\%$). Triage Technologies Inc. offers a tool to detect skin conditions from images that was not a part of the presented experiments.

References

[1] Friedman, R. J., Rigel, D. S. & Kopf, A. W. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians* **35**, 130–151 (1985). URL https://doi.org/10.3322/canjclin.35.3.130.

[2] Henning, J. S. *et al.* The CASH (Color, Architecture, Symmetry, and Homogeneity) Algorithm for Dermoscopy. *Journal of the American Academy of Dermatology* **56**, 45–52 (2007). URL https://linkinghub.elsevier.com/retrieve/pii/S0190962206025278.

[3] Bakheet, S. An SVM Framework for Malignant Melanoma Detection Based on Optimized HOG Features. *Computation* **5**, 4 (2017). URL http://www.mdpi.com/2079-3197/5/1/4.

[4] Grzesiak-Kopeć, K., Nowak, L. & Ogorzałek, M. Automatic Diagnosis of Melanoid Skin Lesions Using Machine Learning Methods. In Rutkowski, L. *et al.* (eds.) *International Conference on Artificial Intelligence and Soft Computing*, 577–585 (Springer, Cham, Zakopane, Poland, 2015). URL http://link.springer.com/10.1007/978-3-319-19324-3{_}51.

[5] Jaworek-Korjakowska, J. Computer-Aided Diagnosis of Micro-Malignant Melanoma Lesions Applying Support Vector Machines. *BioMed Research International* **2016**, 1–8 (2016). URL http://www.hindawi.com/journals/bmri/2016/4381972/.

[6] Murugan, A., Nair, S. H. & Kumar, K. P. S. Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers. *Journal of Medical Systems* **43**, 269 (2019). URL http://link.springer.com/10.1007/s10916-019-1400-8.

[7] Oliveira, R. B., Pereira, A. S. & Tavares, J. M. R. Skin Lesion Computational Diagnosis of Dermoscopic Images: Ensemble Models based on Input Feature Manipulation. *Computer Methods and Programs in Biomedicine* **149**, 43–53 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S0169260717302778.

[8] R D, S. & A, S. Deep Learning Based Skin Lesion Segmentation and Classification of Melanoma Using Support Vector Machine (SVM). *Asian Pacific Journal of Cancer Prevention* **20**, 1555–1561 (2019). URL http://journal.waocp.org/article{_}87402.html.

[9] Esteva, A. *et al.* Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* **542**, 115–118 (2017). URL http://www.nature.com/articles/nature21056.

[10] Haenssle, H. *et al.* Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. *Annals of Oncology* **29**, 1836–1842 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S0923753419341055.

[11] ISIC. ISIC 2019 Skin Lesion Analysis Towards Melanoma Detection (2019). URL https://challenge2019.isic-archive.com/.

[12] Kaggle.com. SIIM-ISIC Melanoma Classification — Kaggle (2020). URL https://www.kaggle.com/c/siim-isic-melanoma-classification/overview.

[13] Hussain, M. A., Amir-Khalili, A., Hamarneh, G. & Abugharbieh, R. Segmentation-Free Kidney Localization and Volume Estimation Using Aggregated Orthogonal Decision CNNs. In Descoteaux, M. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, 612–620 (Springer, Cham, Quebec City, Canada, 2017). URL http://link.springer.com/10.1007/978-3-319-66179-7{_}70.

[14] Lee, H. *et al.* Machine Friendly Machine Learning: Interpretation of Computed Tomography Without Image Reconstruction. *Scientific Reports* **9**, 15540 (2019). URL http://www.nature.com/articles/s41598-019-51779-5.

[15] Taghanaki, S. A. *et al.* Segmentation-free Direct Tumor Volume and Metabolic Activity Estimation from PET Scans. *Computerized Medical Imaging and Graphics* **63**, 52–66 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S0895611117301325.

[16] Xue, W. *et al.* Direct Estimation of Regional Wall Thicknesses via Residual Recurrent Neural Network. In Niethammer, M. *et al.* (eds.) *International Conference on Information Processing in Medical Imaging*, 505–516 (Springer, Cham, Boone, USA, 2017). URL http://link.springer.com/10.1007/978-3-319-59050-9{_}40.

[17] Zhao, R. *et al.* Direct Cup-to-Disc Ratio Estimation for Glaucoma Screening via Semi-Supervised Learning. *IEEE Journal of Biomedical and Health Informatics* **24**, 1104–1113 (2020). URL https://ieeexplore.ieee.org/document/8794624/.

[18] Brinker, T. J. *et al.* Deep Learning Outperformed 136 of 157 Dermatologists in a Head-to-head Dermoscopic Melanoma Image Classification Task. *European Journal of Cancer* **113**, 47–54 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0959804919302217.

[19] Fujisawa, Y. *et al.* Deep Learning Surpasses Dermatologists. *British Journal of Dermatology* **180**, e39–e39 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.17470.

[20] Hekler, A. *et al.* Effects of label noise on deep learning-based skin cancer classification. *Frontiers in Medicine* **7** (2020). URL https://doi.org/10.3389/fmed.2020.00177.

[21] Han, S. S. *et al.* Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Op-

tions for 134 Skin Disorders. *Journal of Investigative Dermatology* (2020). URL https://linkinghub.elsevier.com/retrieve/pii/S0022202X20301366.

[22] Kroemer, S. *et al.* Mobile Teledermatology for Skin Tumour Screening: Diagnostic Accuracy of Clinical and Dermoscopic Image Tele-evaluation Using Cellular Phones. *British Journal of Dermatology* **164**, 973–979 (2011). URL http://doi.wiley.com/10.1111/j.1365-2133.2011.10208.x.

[23] Argenziano, G. *et al.* Interactive Atlas of Dermoscopy: A Tutorial (Book and CD-ROM) (2000). URL http://www.dermoscopy.org/atlas/order{_}cd.asp.

[24] Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics* **23**, 538–546 (2019). URL https://ieeexplore.ieee.org/document/8333693/.

[25] Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. 7-Point Criteria Evaluation Database (2019). URL https://derm.cs.sfu.ca/.

[26] Brinker, T. J. *et al.* Comparing artificial intelligence algorithms to 157 german dermatologists: the melanoma classification benchmark. *European Journal of Cancer* **111**, 30–37 (2019). URL https://doi.org/10.1016/j.ejca.2018.12.016.

[27] Zamir, A. R. *et al.* Taskonomy: Disentangling Task Transfer Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3712–3722 (IEEE, Salt Lake City, USA, 2018). URL https://ieeexplore.ieee.org/document/8578489/.

[28] Argenziano, G. *et al.* Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions. *Archives of Dermatology* **134** (1998). URL http://archderm.jamanetwork.com/article.aspx?doi=10.1001/archderm.134.12.1563.

[29] Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 233–240 (ACM Press, Pittsburgh, USA, 2006). URL http://portal.acm.org/citation.cfm?doid=1143844.1143874.

[30] Yoon, C., Hamarneh, G. & Garbi, R. Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, 365–373 (Springer International Publishing, 2019). URL https://doi.org/10.1007/978-3-030-32251-9_40.

[31] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal Trial of an Autonomous AI-based Diagnostic System for Detection of Diabetic Retinopathy in Primary

Care Offices. *NPJ Digital Medicine* **1**, 39 (2018). URL http://www.nature.com/articles/s41746-018-0040-6.

[32] BusinessWire. Zebra Medical Vision Secures a Fourth FDA Clearance for AI for Medical Imaging (2019). URL https://www.businesswire.com/news/home/20191127005391/en/Zebra-Medical-Vision-Secures-Fourth-FDA-Clearance.

[33] FDA. FDA Permits Marketing of Artificial Intelligence-based Device to Detect Certain Diabetes-related Eye Problems (2018). URL https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

[34] Carli, P. *et al.* Pattern Analysis, Not Simplified Algorithms, is the Most Reliable Method for Teaching Dermoscopy for Melanoma Diagnosis to Residents in Dermatology. *British Journal of Dermatology* **148**, 981–984 (2003). URL http://doi.wiley.com/10.1046/j.1365-2133.2003.05023.x.

[35] Johr, R. H. Interactive CD of Dermascopy. *Archives of Dermatology* **137**, 831–832 (2001).

[36] Lio, P. A. & Nghiem, P. Interactive Atlas of Dermoscopy. *Journal of the American Academy of Dermatology* **50**, 807–808 (2004). URL https://linkinghub.elsevier.com/retrieve/pii/S0190962203033358.

[37] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826 (IEEE, Las Vegas, USA, 2016). URL http://ieeexplore.ieee.org/document/7780677/.

[38] Deng, J. *et al.* ImageNet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, Miami, USA, 2009). URL https://ieeexplore.ieee.org/document/5206848/.

[39] Vesal, S., Patil, S. M., Ravikumar, N. & Maier, A. K. A multi-task framework for skin lesion detection and segmentation. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, 285–293 (Springer International Publishing, 2018). URL https://doi.org/10.1007/978-3-030-01201-4_31.

[40] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**, 80 (1945). URL https://doi.org/10.2307/3001968.

[41] Efron, B. Bootstrap Methods: Another Look at the Jackknife. In Kotz, S. & Johnson, N. L. (eds.) *Springer Series in Statistics (Perspectives in Statistics)*, 569–593 (Springer, New York, NY, 1992). URL http://link.springer.com/10.1007/978-1-4612-4380-9{_}41.

[42] Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R., Moret, B. M. & Stamatakis, A. How Many Bootstrap Replicates Are Necessary? *Journal of Computational Biology* **17**, 337–354 (2010). URL http://www.liebertpub.com/doi/10.1089/cmb.2009.0179.

[43] Athreya, K. B. Bootstrap of the Mean in the Infinite Variance Case. *The Annals of Statistics* **15**, 724–731 (1987). URL http://projecteuclid.org/euclid.aos/1176350371.

[44] Cohen, J. *Statistical power analysis for the behavioral sciences* (Academic Press, 1977).

[45] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46 (1960). URL https://doi.org/10.1177/001316446002000104.

[46] Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (1971). URL https://doi.org/10.1037/h0031619.

[47] van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13**, 22–30 (2011). URL http://ieeexplore.ieee.org/document/5725236/.

[48] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020). URL http://www.nature.com/articles/s41592-019-0686-2.

[49] Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* (2010).

[50] Haghighi, S., Jasemi, M., Hessabi, S. & Zolanvari, A. PyCM: Multiclass confusion matrix library in python. *Journal of Open Source Software* **3**, 729 (2018). URL https://doi.org/10.21105/joss.00729.

[51] Pedregosa, F. *et al.* Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[52] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007). URL http://ieeexplore.ieee.org/document/4160265/.

[53] Waskom, M. *et al.* mwaskom/seaborn: v0.8.1 (September 2017) (2017). URL https://doi.org/10.5281/zenodo.883859.

[54] Chollet, F. *et al.* Keras. https://keras.io (2015).

Figure Legends

Figure 1: An overview of the three prediction models. All the models take the clinical and the dermoscopic images of the skin lesion and the patient metadata as input. (a) The first model predicts the lesion diagnosis probabilities, $\text{DIAG}_{\text{pred}}$. (b) The second model predicts the management decision probabilities, $\text{MGMT}_{\text{pred}}$. (c) The third is a multi-task model and predicts the seven-point criteria ($\text{Criterion}\{1, 2, \cdots, 7\}_{\text{pred, multi}}$) in addition to $\text{DIAG}_{\text{pred, multi}}$ and $\text{MGMT}_{\text{pred, multi}}$. The argmax operation assigns 1 to the most likely label and 0 to all others. For (a), $\text{DIAG}_{\text{pred}}$ diagnosis is used to arrive at a management decision either using (a1) binary labeling, $\text{MGMT}_{\text{infr, binary}}$, or (a2) prior based inference, $\text{MGMT}_{\text{infr, all}}$. Similarly, the outputs of (b) can be used to directly predict a management decision using either (b1) binary labeling, $\text{MGMT}_{\text{pred, binary}}$, or (b2) all the labels, $\text{MGMT}_{\text{pred, all}}$. As explained in the text, the diagnosis labels are basal cell carcinoma (BCC), nevus (NEV), melanoma (MEL), seborrheic keratosis (SK), and others (MISC), and the management decision labels are 'clinical follow up' (CLNC), 'excision' (EXC), and 'no further examination' (NONE). In the case of binary management decisions, we predict whether a lesion should be excised (EXC) or not (NOEXC).

Figure 2: Quantitative evaluation of the $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ predictions. (a) Violin plots of the distance measures of the probabilistic predictions show that the $\text{MGMT}_{\text{pred, all}}$ predictions are closer (statistically significant) to the target labels for test data. (b, c) ROC curves and (d, e) confusion matrices of $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$ respectively along with cell-wise diagnosis breakdown. Note that $\text{MGMT}_{\text{infr, all}}$ has a tendency to over-excise lesions.

Figure 3: Evaluating the multi-modal multi-task model. (a) ROC curve and (b) precision-recall curve for the management prediction task. Confusion matrices for (c) the management prediction task and (d) the diagnosis prediction task along with the diagnosis-wise breakdown for the management labels.

Figure 4: Evaluating the multi-task model on the MClass-D dataset. (a) Confusion matrices and (b) ROC curves for $\text{MGMT}_{\text{pred}}$ and $\text{MGMT}_{\text{infr}}$ predictions with both $\text{MGMT}_{\text{GT, agg}}$ and $\text{MGMT}_{\text{GT, true}}$ as target clinical management labels.

Figure 5: A breakdown of the inputs, outputs, loss functions, and architecture of the three prediction models. Global average pooled feature responses from the clinical and the dermoscopic images are extracted and concatenated (denoted by the plus symbol) with one-hot encoded patient meta-data, and the three models are trained with L_{DIAG}, L_{MGMT}, and L_{multi} respectively. The first model predicts the diagnosis labels ($\text{DIAG}_{\text{pred}}$) which are then used along with the management priors to obtain inferred management decisions ($\text{MGMT}_{\text{infr}}$), whereas the second model predicts the management decisions directly ($\text{MGMT}_{\text{pred}}$). Finally, the last model is a

multi-task one and is trained to predict the seven-point criteria, the diagnosis, and the management (outputs enclosed in the dashed box).

Tables

Table 1: Breakdown of the seven-point criteria evaluation dataset [25] by management and diagnosis labels and the train-valid-test splits used to train the model.

Management	**Diagnosis**					**Split**			**Total**
	BCC	NEV	MEL	MISC	SK	Training	Validation	Testing	
CLNC	0	268	0	24	4	133	51	112	296
EXC	42	278	252	23	10	235	127	243	605
NONE	0	29	0	50	31	45	25	40	110
Total	42	575	252	97	45	413	203	395	

Table 2: Comparing skin lesion management prediction results obtained using $\text{MGMT}_{\text{infr, all}}$ and $\text{MGMT}_{\text{pred, all}}$.

Management	**$\text{MGMT}_{\text{infr, all}}$**				**$\text{MGMT}_{\text{pred, all}}$**			
Labels	Sensitivity	Specificity	Precision	AUROC	Sensitivity	Specificity	Precision	AUROC
NONE	0.2	0.9718	0.4444	0.8039	0.5	0.9831	0.7692	0.9159
CLNC	0.0	1.0	0.0	0.7668	0.7143	0.7456	0.5263	0.8090
EXC	0.9835	0.0921	0.634	0.7515	0.7243	0.7303	0.8111	0.8079
Average	0.3945	0.6880	0.3595	0.7741	**0.6462**	**0.8196**	**0.7022**	**0.8443**

Table 3: Skin lesion management prediction results using a multi-modal multi-task model.

Management Label	**Metrics**			
	Sensitivity	Specificity	Precision	AUROC
NONE	0.6500	0.9747	0.7429	0.9225
CLNC	0.6071	0.8375	0.5965	0.8065
EXC	0.8107	0.6776	0.8008	0.8226
Average	0.6893	0.8299	0.7134	0.8505