

1 **Deep learning for AI-based diagnosis of skin-related neglected tropical**
2 **diseases: a pilot study**

3
4

5 Rie Yotsu¹, Zhengming Ding², Jihun Hamm², Ronald Blanton¹

6 ¹Department of Tropical Medicine, Tulane University School of Public Health and
7 Tropical Medicine, 1440 Canal Street, New Orleans, LA 70112

8 ²Department of Computer Science, Tulane University School of Science and
9 Engineering, 201 Lindy Claiborne Boggs Center, 6823 St. Charles Avenue, New
10 Orleans, LA 70118

11
12

13 Corresponding author: Rie Yotsu, Tulane University School of Public Health and
14 Tropical Medicine, ryotsu@tulane.edu

15
16

17 **ABSTRACT**

18 **Background** Deep learning, which is a part of a broader concept of artificial
19 intelligence (AI) and/or machine learning has achieved remarkable success in
20 vision tasks. While there is growing interest in the use of this technology in
21 diagnostic support for skin-related neglected tropical diseases (skin NTDs), there
22 have been limited studies in this area and fewer focused on dark skin. In this
23 study, we aimed to develop deep learning based AI models with clinical images
24 we collected for five skin NTDs, namely, Buruli ulcer, leprosy, mycetoma, scabies,
25 and yaws, to understand how diagnostic accuracy can or cannot be improved
26 using different models and training patterns.

27 **Methodology** This study used photographs collected prospectively in Côte
28 d'Ivoire and Ghana through our ongoing studies with use of digital health tools for
29 clinical data documentation and for teledermatology. Our dataset included a total
30 of 1,709 images from 506 patients. Two convolutional neural networks, ResNet-
31 50 and VGG-16 models were adopted to examine the performance of different
32 deep learning architectures and validate their feasibility in diagnosis of the
33 targeted skin NTDs.

34 **Principal findings** The two models were able to correctly predict over 70% of
35 the diagnoses, and there was a consistent performance improvement with more
36 training samples. The ResNet-50 model performed better than the VGG-16
37 model. Model trained with PCR confirmed cases of Buruli ulcer yielded 1-3%
38 increase in prediction accuracy over training sets including unconfirmed cases.

39 **Conclusions** Our approach was to have the deep learning model distinguish
40 between multiple pathologies simultaneously – which is close to real-world
41 practice. The more images used for training, the more accurate the diagnosis

42 became. The percentages of correct diagnosis increased with PCR-positive
43 cases of Buruli ulcer. This demonstrated that it may be better to input images
44 from the more accurately diagnosed cases in the training models also for
45 achieving better accuracy in the generated AI models. However, the increase was
46 marginal which may be an indication that the accuracy of clinical diagnosis alone
47 is reliable to an extent for Buruli ulcer. Diagnostic tests also have its flaws, and
48 they are not always reliable. One hope for AI is that it will objectively resolve this
49 gap between diagnostic tests and clinical diagnoses with addition of another tool.
50 While there are still challenges to be overcome, there is a potential for AI to
51 address the unmet needs where access to medical care is limited, like for those
52 affected by skin NTDs.

53

54 **Keywords**

55 artificial intelligence; Buruli ulcer; deep learning; Hansen's disease; leprosy;
56 machine learning; mycetoma; scabies; skin; skin NTDs; skin of color; yaws

57

58

59 **AUTHOR SUMMARY**

60 The diagnosis of skin diseases depends in large part, though not exclusively on
61 visual inspection. The diagnosis and management of these diseases is thus
62 particularly amenable to teledermatology approaches. The widespread
63 availability of cell phone technology and electronic information transfer provides
64 new potential for access to health care in low-income countries, yet there are
65 limited efforts targeting these neglected populations with dark skin and
66 consequently limited availability of tools. In this study, we leveraged a collection
67 of skin images gathered through a system of teledermatology in the West African
68 countries of Côte d'Ivoire and Ghana, and applied deep learning, a form of
69 artificial intelligence (AI) - to see if deep learning models can distinguish between
70 different diseases and support their diagnosis. Skin-related neglected tropical
71 diseases, or skin NTDs, prevail in these regions and were our target conditions:
72 Buruli ulcer, leprosy, mycetoma, scabies, and yaws. The accuracy of prediction
73 depended on the number of images that were fed into the model for training with
74 marginal improvement using laboratory confirmed cases in training. Using more
75 images and greater efforts in this area, it is possible that AI can help address the
76 unmet needs where access to medical care is limited.

77

78

79 INTRODUCTION

80 Deep learning has achieved remarkable success in vision tasks such as
81 image classification, image localization and image semantic segmentation, which
82 also includes skin disease prediction. Deep learning is a part of a broader concept
83 of artificial intelligence and/or machine learning whereby it uses vast volumes of
84 data and complex algorithms to train a model to perform certain tasks. The
85 success of the approach undoubtedly can be attributed to the ability of learning
86 abstract semantic knowledge with the hierarchical network architecture from
87 visual signals (1). It is increasingly gaining interest and becoming more important
88 in the field of dermatology in this digital era. Evidence is accumulating that deep
89 learning can assist healthcare providers to make better clinical decisions, even
90 to an extent that sometimes exceeds human judgement (2, 3, 4). However, many
91 of the diseases studied are pigmented lesions such as melanoma and basal cell
92 carcinoma, or inflammatory dermatoses which often affect people with lighter skin
93 color and thus provide a high degree of contrast (5, 6, 7).

94 Skin-related neglected tropical diseases, or skin NTDs, comprise a group
95 of infectious diseases whose morbidity is expressed on the skin. They include at
96 least nine diseases and disease groups listed by the World Health Organization
97 (WHO) (8). More than 1 billion people are known to be either at risk or infected
98 by skin NTDs (9). They mainly prevail in poor communities of low- and middle-
99 income countries (LMICs) where resources are scarce and where there are
100 limited numbers of dermatologists to diagnose the conditions. Additionally, skin
101 NTDs more often affect people of color. Availability of screening systems,
102 therefore, is critical for this set of diseases which will enable earlier diagnosis and

103 treatment. The longer the delay in diagnosis, the more patients with skin NTDs
104 may be left with life-long disabilities and deformities.

105 While there is growing interest in the use of deep learning for diagnosis
106 of skin NTDs to fill in these gaps, there have been limited studies to date which
107 investigated the development of an AI model for a combination of these less
108 studied diseases in the less studied populations. In this study, we aimed to
109 develop deep learning based AI models with clinical images we collected for five
110 skin NTDs, namely, Buruli ulcer, leprosy, mycetoma, scabies, and yaws, to
111 understand how diagnostic accuracy is influenced by different models, especially
112 when the training images are relatively small in number and collected under
113 diverse conditions. All of the images are from dark-skinned African populations,
114 with Fitzpatrick skin type IV or above. We anticipate that our findings will support
115 future development of AI models for the skin NTDs, and in addition, other skin
116 diseases in people with darker skin types.

117

118 **METHODS**

119 This study used photographs that were collected prospectively in the
120 West African countries of Côte d'Ivoire and Ghana, through our ongoing studies
121 with use of digital health tools for clinical data documentation and for
122 teledermatology. The description of the design of this study can be found
123 elsewhere (10). Briefly, photographs of skin lesions were collected with clinical
124 information including demographics and disease description to support
125 dermatologists in providing diagnoses remotely. The photographs were taken
126 using the camera on Lenovo Tab M10 FHD Plus smart tablets under field
127 conditions and in rural clinics from a total of six health districts (four in Côte

128 d'Ivoire and two in Ghana) known to be endemic with one or more skin NTDs.
129 Image resolution was 1920 x 2560 pixels stored in JPEG format. Written informed
130 consent was obtained from all patients for use of their images. The study has
131 ethical approvals from the institutional review board of the Tulane University
132 School of Public Health and Tropical Medicine (2020-2054-SPHTM) (USA), the
133 Ministry of Health of Côte d'Ivoire (No. IRB000111917), and the Ministry of Health
134 of Ghana (GHS-ERC:014/05/21).

135

136 **Dataset screening**

137 Images were selected from our data repository for which diagnoses for
138 one of the five targeted diseases (Buruli ulcer, leprosy, mycetoma, scabies, and
139 yaws) were made remotely and in person by two dermatologists with more than
140 10 years of experience in diagnosing patients locally. A portion of cases of Buruli
141 ulcer underwent polymerase chain reaction (PCR) testing for confirmation.
142 Likewise, dual path platform (treponemal and non-treponemal) (DPP) testing
143 (Chembio Diagnostics, Medford, NY, USA) was done for a portion of cases of
144 yaws. Table 1 shows the data summary of the five diseases, with number of
145 patients and number of images for each disease. Multiple images were obtained
146 for most patients. For Buruli ulcer and yaws, the numbers within parenthesis were
147 those with positive results with PCR and DPP, respectively.

148

149 **Table 1.** Patient and image sample sizes

| | Buruli ulcer* | Leprosy | Mycetoma | Scabies | Yaws* |
|----------|---------------|---------|----------|---------|-----------|
| Patients | 200 (97) | 38 | 12 | 107 | 149 (62) |
| Images | 784 (361) | 131 | 32 | 389 | 373 (147) |

150 *Parentheses indicate the number with positive PCR or DPP test results for Buruli
151 ulcer or yaws, respectively. Multiple photographs were taken for each patient.

152

153 **AI-based skin disease diagnosis model**

154 Convolutional neural networks (CNNs) are the popular deep learning
155 techniques to extract feature representation from the image samples for disease
156 diagnosis. CNNs are multi-layer neural networks with convolutional filters to
157 capture the visual pattern from skin images. In this study, we adopted two popular
158 CNNs, the ResNet-50 (50-layer residual neural network) (11) and the 16-layer
159 VGG-16 model (12). The purpose was to examine the performance of different
160 deep learning architectures and validate the feasibility of deep learning models
161 in diagnosis of these skin diseases.

162 All the original images were resized into the same size as a 3D tensor
163 with 224 x 224 x 3 pixel resolution to fit the input of deep learning models. Data
164 augmentations and normalization pre-processing strategies were also employed
165 following existing image classification tasks (13). These were then sent to the
166 Resnet-50 model or VGG-16 model, pre-trained on the ImageNet dataset, which
167 is a large-scale, open-source image repository (14). Each image was represented
168 as a 2048-dimensional feature vector for ResNet-50, and a 4096-dimensional
169 feature vector for VGG-16. Following this, we designed the disease diagnosis
170 classifier with output of a 5-dimensional vector as a 5-disease probability vector.
171 For model optimization, we adopted stochastic gradient descent (SGD) with a
172 momentum of 0.9 as optimizer to update the whole network parameters (*i.e.*,
173 ResNet50 and classifier parameters). We performed the experiments using the
174 PyTorch library running on one GPU (NVIDIA Titan V).

175 To train our designed models, the images from k% of patients from our
176 collection chosen at random was used as a training set, and the images from the
177 remaining patients were used to evaluate the model performance. We tested if
178 and how laboratory confirmation may change the accuracy of the classifier. With
179 our datasets (Table 1), we performed two kinds of experiments: firstly, using all
180 cases [clinical diagnosis] (Task 1), and secondly, using only those cases that
181 tested positive with PCR or DPP for Buruli ulcer and yaws, respectively [test
182 positives] (Task 2). Otherwise, the analysis was the same. There was no patient
183 overlap between the training and test sets.

184 We adopted two metrics, the Top-1 accuracy (%) and the Matthew's
185 correlation coefficient (MCC, 0~1) to evaluate our model (15). Top-1 accuracy
186 measures the proportion of test images for which the predicted disease matches
187 the single target disease. MCC is a reliable statistical score that produces a high
188 value only if the prediction obtained good results in all the four confusion matrix
189 categories (true positives, false negatives, true negatives, and false positives).
190 To map the learned visual representations, we used the dimensionality reduction
191 method, Principal Component Analysis (PCA) (16).

192

193 **RESULTS**

194 Table 2 presents the results of diagnostic accuracy from the two models
195 (ResNet-50 and VGG-16) using all images as Task 1 and using images from
196 laboratory confirmed positive cases of Buruli ulcer and yaws in training as Task
197 2. From the results across Tasks 1 and 2, we observed a consistent performance
198 improvement when we had more training samples.

199

200 **Table 2:** Diagnostic accuracy of the two network models with different
201 percentages of images for training

202 (a) Task 1: Clinical diagnosis

203 ResNet-50

| k% | 30 | 40 | 50 | 60 | 70 |
|-----------|--------|--------|--------|--------|--------|
| Top-1 Acc | 78.55% | 79.36% | 82.20% | 82.84% | 84.63% |
| MCC | 0.6776 | 0.6897 | 0.7308 | 0.7394 | 0.7715 |

204 VGG-16

| k% | 30 | 40 | 50 | 60 | 70 |
|-----------|--------|--------|--------|--------|--------|
| Top-1 Acc | 76.61% | 78.45% | 80.55% | 80.27% | 82.22% |
| MCC | 0.6440 | 0.6758 | 0.7038 | 0.6984 | 0.7336 |

205

206 (b) Task 2: Lab test positives

207 ResNet-50

| k% | 30 | 40 | 50 | 60 | 70 |
|-----------|--------|--------|--------|--------|--------|
| Top-1 Acc | 75.71% | 76.52% | 77.19% | 79.15% | 84.17% |
| MCC | 0.6505 | 0.6615 | 0.6725 | 0.7011 | 0.7706 |

208

209 VGG-16

| k% | 30 | 40 | 50 | 60 | 70 |
|-----------|--------|--------|---------|--------|--------|
| Top-1 Acc | 73.64% | 74.93% | 75.79%% | 78.72% | 79.44% |
| MCC | 0.6204 | 0.6407 | 0.6515 | 0.6948 | 0.7011 |

210 ResNet-50 and VGG-16 are two CNN image models with the former having the higher number
211 of layers. K% is the percentage of all images used in training. Training and test samples did not
212 overlap. Top-1 Acc: top-1 accuracy is the proportion of test images for which the predicted
213 diagnosis by the model matches the actual diagnosis. MCC: Matthew's correlation coefficient is
214 a statistical value assessed by results obtained from four confusion matrix categories. A higher
215 value means better prediction (range, 0-1).

216

217 ***Confirmatory analysis and confusion matrix***

218 To further understand our AI diagnosis model for each disease, we analyzed the
219 confusion matrices of the two models to examine if the model trained with images
220 from the diagnostic test positive data would contribute to better performance. For
221 Setting A, we used 50 PCR positive Buruli ulcer patients with PCR positives plus
222 the other 4 diseases to train the model based on ResNet-50. For Setting B, we
223 used 50 clinically diagnosed Buruli ulcer patients (including every patient
224 diagnosed as Buruli ulcer irrespective of their PCR results) plus the other 4
225 diseases to train the model based on ResNet-50. The test set was the same for
226 the two models, which included 100 clinically diagnosed Buruli ulcer patients plus
227 the other 4 diseases. There was no patient overlap between the training and test
228 set. Overall, Setting A was able to achieve 80% accuracy while this was 77% for
229 Setting B.

230 Figure 1 shows confusion matrices by each disease. Each row of the
231 matrix represents the instances of actual diagnosis [ground truth], while each
232 column represents the instances of predicted diagnosis by the deep learning
233 model [prediction]. Each diagonal element denotes the correct diagnosis by the
234 model per disease. The prediction accuracy increased in Setting A by 1-3% as
235 compared to Setting B across all diseases besides for mycetoma, where there
236 were smaller number of images. For both Settings A and B, Buruli ulcer and
237 scabies had the highest percentages of correct diagnosis, 88% and 85% for both,
238 respectively.

239

240 ***Qualitative Analysis***

241 Figure 2 provides eight example images which resulted in incorrect
242 prediction by our pilot AI model based on ResNet-50, with (k=50)% training data
243 for all data (Task 1). Numbers in the parentheses represent the likelihood of the
244 diagnosis by the prediction model [prediction label] as compared to the actual
245 diagnosis [true label], or the ground truth. An uncertainty score is also given to
246 each test image, which is calculated by the correlation between the predicted
247 probability with random guess. Higher correlation means higher uncertainty score.
248 The uncertainty score indicates the degree of irrelevant evidence the AI model
249 finds for the given test image used to predict its diagnosis. For example, Figure
250 2(a) shows a true label score for yaws of 0.187 and a predicted label for Buruli
251 ulcer of 0.254 with high uncertainty of 0.93. This means that the model predicted
252 the image to be more like Buruli ulcer than yaws, however it was also highly
253 uncertain. An uncertainty score closer to 1 represents higher uncertainty for the
254 diagnosis output. When it is 100% uncertain, AI estimates it to be a random guess
255 and provides a confidence score of 0.200 (5 diseases, $1/5 = 0.200$). The AI
256 prediction is better when the uncertainty score is lower, although the diagnosis
257 could still be incorrect.

258

259 ***Feature Visualization***

260 To further understand why we can achieve better performance on Buruli
261 ulcer and scabies but worse performance for instance on mycetoma, we used,
262 PCA (16) to map the learned visual representations (2048-dimensional features
263 of ResNet-50) of each test class image to a 2-D plane. The goal was to visualize
264 the learned feature representation and provide a direct way to understand the
265 discriminative ability of AI features from raw skin images. Figure 3a shows the

266 training samples while Figure 3b lists the test samples, where each dot in the 2-
267 D plane denotes one image sample and the same color means the image
268 samples of the same disease. From the results, we observed that our
269 classification model can learn discriminative features from raw skin images to
270 differentiate diseases in the training stage, while the model generalization ability
271 to test images becomes poorer, which means the model cannot easily
272 differentiate the test images like the training ones.

273

274 **DISCUSSION**

275 In this report, we explored how deep learning might help in screening
276 and/or diagnosis of skin NTDs, which often affects people with darker skin tones.
277 Two deep learning models were examined in our work. Between the ResNet-50
278 and VGG-16 models, we conclude that the ResNet-50 model achieved better
279 performance (around 2% better prediction for all evaluation) in predicting our skin
280 images. The major difference between the two models is the depth of their layers,
281 *i.e.*, ResNet-50 contains 50 layers of convolutional, pooling operations, while
282 VGG-16 only contains 16 layers of the same. Generally, deeper models with more
283 layers can extract more powerful representations from image data (12). This
284 tendency was consistent also for our dataset which focused on skin disease
285 diagnosis. However, models with more layers contain more parameters, which
286 make them heavier, and less efficient (12). VGG-16 is more efficient as fewer
287 layers are included.

288 Although classified together as skin NTDs, the target infections have
289 quite different appearances, presentations, and progressions. The lesions can be
290 raised, depressed, smooth or rough, various colors or multicolored even for the

291 same condition. We observed that deep learning approaches to identification of
292 Buruli ulcer, scabies and yaws showed good performance of close to over 80%
293 prediction, perhaps since these were trained with more images. Especially for
294 Buruli ulcer where there were 784 images, the performance was over 90%.
295 Leprosy and mycetoma, used smaller sample sizes and had poorer performance.
296 For leprosy, we speculate that it was not only the sample size, but also the
297 complexity of the disease presentation that impacted performance (17). We had
298 a range of images from tuberculoid to borderline to lepromatous type leprosy, as
299 well as some included deformities and wounds that developed due to peripheral
300 neuropathy. We stratified these different conditions and ran the same analysis,
301 with an expectation that this may increase power by reducing variance. However,
302 this further decreased the number of our samples, and we were unable to obtain
303 any meaningful results this time. Likewise, similar results were obtained for yaws,
304 when stratified for ulcerative versus non-ulcerative (papilloma, hyperkeratosis,
305 etc.) lesions. However, we believe if there were enough images, stratifying may
306 increase the accuracy of the predicted diagnosis. Moreover, as prior study on AI-
307 based diagnosis for leprosy showed, clinical data other than images, most
308 importantly loss of sensation for leprosy, are essential to be combined in the deep
309 learning dataset for better model development (17).

310 The percentages of correct diagnosis increased with PCR-positive cases
311 of Buruli ulcer. This demonstrated that it may be better to input images from the
312 more accurately diagnosed cases in the training models also for achieving better
313 accuracy in the generated AI models. It was interesting to see that the PCR-
314 confirmed cases of Buruli ulcer contributed in increasing the diagnostic accuracy
315 not just for Buruli ulcer but also for other diseases. On the other hand, contrary

316 to our hypothesis, the percentage increase was minimal (3% for Buruli ulcer),
317 which may be an indication that the accuracy of clinical diagnosis alone is reliable
318 to an extent. Especially for Buruli ulcer, a previous study by Eddyani et al. has
319 shown that sensitivity of clinical diagnosis was as high as 92% (95% CI, 85-96%),
320 which was the highest among any other methods including PCR (18). PCR results
321 can be false negatives in Buruli ulcer due to several factors, for example, site of
322 sample collection, skills in sample taking and duration of the wound (19). While it
323 is currently the preferred test for diagnostic confirmation, it has its flaws and is
324 not always reliable. In many studies, PCR is considered 65-70% sensitive (20) or
325 even only 61% sensitive (21). Specificity is perhaps highest for the PCR positive
326 cases, but sensitivity is highest for clinically identified cases. The PCR positive
327 cases should be enriched for true cases, but it also misses true cases. One hope
328 for AI – which our findings also support – is that it will objectively resolve this gap
329 between diagnostic tests like PCR and clinical diagnoses with addition of another
330 tool.

331 Incorrect diagnoses made by our model were skewed towards other skin
332 NTDs being diagnosed as Buruli ulcer, as about half of our images were Buruli
333 ulcer. Fairness issues in deep learning arise when the dataset is extremely
334 imbalanced across different categories or groups (22). When these images with
335 incorrect prediction were reviewed, some cases would have been difficult to
336 differentiate even with the human eye, such as the case of yaws in Figure 2(b),
337 for example. On the other hand, some cases with obviously different
338 presentations were predicted to be Buruli ulcer, such as the case of mycetoma in
339 (d) and leprosy in (f). We were unsure why they were predicted to be Buruli ulcer.
340 For cases shown in (a) and (g), images and location on the limbs may have

341 played some role in these being predicted as a Buruli ulcer case, as the most
342 commonly affect body parts in Buruli ulcer are the limbs (23, 24). Figure 2(c) was
343 a case of yaws, but the main lesion was not centered, and the lesion of interest
344 was not very obvious. The backgrounds or the clothes may have disturbed the
345 predictions in cases such as in (a), (e), (g), and (h). It will be necessary to
346 understand these patterns in order to resolve incorrect predictions, which will be
347 one of our future study directions.

348 A major source of bias in AI applications stems from the availability and
349 variety of images used in training. There are a very limited number of images of
350 these diseases and a more limited number of images of people of color. In
351 addition, the phrase "people of color" embraces a huge range of hues and surface
352 characteristics even within the African continent. One of the strong points of this
353 pilot has been the use of local dermatologists. In one example in the field, the
354 local dermatologists recognized a series of deeply pigmented lesions as being a
355 reaction to skin whitening agents, a diagnosis that would not easily be arrived at
356 by physicians in the US or Europe. A key observation here has been to reinforce
357 the need for more images from a wider diversity of cases from this part of the
358 world, similarly to the recently recognized gap in dermatological training in
359 general (25, 26). We were able to derive almost the same, if not close, accuracy
360 in diagnosis with the model trained with images from clinical diagnosis over those
361 trained with images with laboratory confirmation – this was partly possible
362 because of the involvement of our skilled local dermatologists.

363 There are limitations to our study, some of which were already described,
364 such as limited number of images and imbalance in image numbers between
365 diseases. Moreover, images were taken under different conditions, and they were

366 highly heterogeneous, for example, distracting objects in the background or
367 lighting. We are currently working on how to mitigate them, as the photos are
368 taken under field conditions in Côte d'Ivoire and Ghana where conditions are less
369 formal than with many other studies. As it is difficult to mandate the images be
370 taken in a uniform environment in these settings, and as this will also limit the
371 number of images that we can use for deep learning, it is potentially more up to
372 the technology how we can overcome this challenge. If such technology can be
373 developed, it could be beneficial for the development of deep learning models for
374 a wide range of skin diseases that are common in the developing world, in
375 addition to what were targeted in this study.

376

377 **Conclusions**

378 Here, we presented our exploratory approach in developing deep
379 learning models for skin NTDs and the challenges that we encountered. These
380 attempts have only just begun. We hope that the lessons learnt here will support
381 the future development of AI technology for these neglected diseases in the
382 neglected populations. Our approach was to have the deep learning model
383 distinguish between multiple pathologies simultaneously. This is different from
384 many other studies where deep learning models were asked to make a diagnosis
385 of a single disease. However, in real-world, what happens in clinicians' mind is
386 that we are required to compare between different pathologies – accordingly, we
387 devised an approach that is more in line with this practice. AI is not yet a
388 replacement for human diagnosis, but if used well and appropriately, it is a tool
389 that can be useful in screening for diseases and improving patient outcomes.

390 Particularly, the hope is that it will address the unmet needs where access to
391 medical care is limited, like for those affected by skin NTDs.
392

393 **Declaration of interests**

394 All authors declare no competing conflict of interests.

395

396 **Funding**

397 Research reported in this publication is supported by the Fogarty International

398 Center of the National Institutes of Health under award number R21TW011860.

399 The content is solely the responsibility of the authors and does not necessarily

400 represent the official views of the National Institutes of Health. Additional

401 funding for this study comes from the International Collaborative Research

402 Program for Tackling the NTDs (Neglected Tropical Diseases) Challenges in

403 African countries by the Japan Agency for Medical Research and Development

404 (AMED) under award grant number 21jm0510004h0204, the Leprosy Research

405 Initiative (LRI) under grant number 707.19.62, and the Global Health Innovative

406 Technology (GHIT) under award grant number G2020-202.

407

408 **Data availability**

409 The data sets generated or analyzed during this study are available from the

410 corresponding author on reasonable request.

411

412 **Acknowledgements**

413 We would like to pay special thanks to Prof. Bamba Vagamon and Prof. Almamy
414 Diabate (Université Alassane Ouattara, Service de Dermatologie CHU de
415 Bouaké-Côte d'Ivoire) for their support in diagnosis of skin diseases on site and
416 remotely. We would like to also thank the project team, Aubin Yao and Luc
417 Kowaci Gontran Yeboue (Hope Commission International) for their hard work in
418 implementing the teledermatology project.

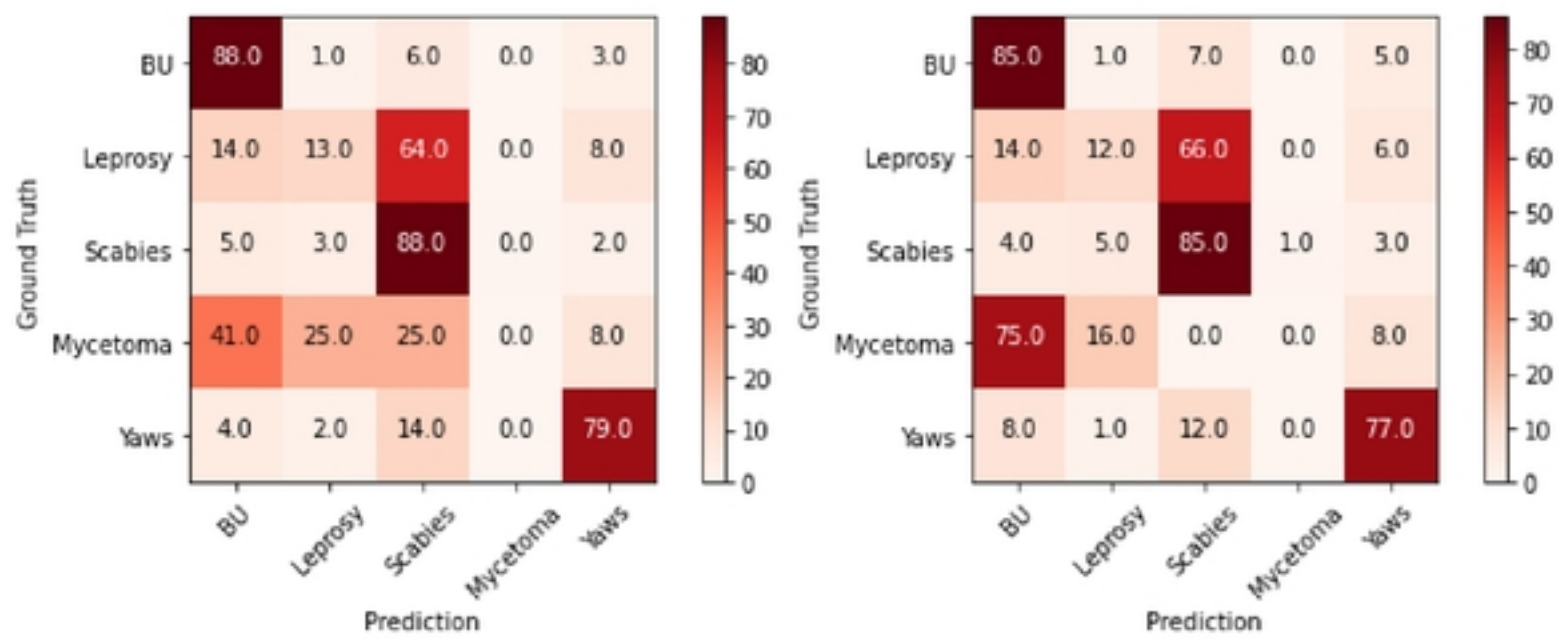
419

420 **References**

- 421 1. Patel S, Wang JV, Motaparathi K, Lee JB. Artificial Intelligence in dermatology
422 for the clinicians. *Clinics in Dermatology*. 2021;39:667-72.
- 423 2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al.
424 Dermatologist-level classification of skin cancer with deep neural networks.
425 *Nature*. 2017;542(7639):115-8.
- 426 3. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al.
427 Man against machine: diagnostic performance of a deep learning convolutional
428 neural network for dermoscopic melanoma recognition in comparison to 58
429 dermatologists. *Ann Oncol*. 2018;29(8):1836-42.
- 430 4. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al.
431 Deep neural networks are superior to dermatologists in melanoma image
432 classification. *Eur J Cancer*. 2019;119:11-7.
- 433 5. Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K,
434 Ranmuthu CKI, Islam MS, et al. Artificial intelligence and machine learning
435 algorithms for early detection of skin cancer in community and primary care
436 settings: a systematic review. *Lancet Digit Health*. 2022;4(6):e466-e76.
- 437 6. Dick V, Sinz C, Mittlbock M, Kittler H, Tschandl P. Accuracy of Computer-
438 Aided Diagnosis of Melanoma: A Meta-analysis. *JAMA Dermatol*.
439 2019;155(11):1291-9.
- 440 7. Seite S, Khammari A, Benzaquen M, Moyal D, Dreno B. Development and
441 accuracy of an artificial intelligence algorithm for acne grading from smartphone
442 photographs. *Exp Dermatol*. 2019;28(11):1252-7.
- 443 8. World Health Organization. Ending the neglect to attain the Sustainable
444 Development Goals: a strategic framework for integrated control and
445 management of skin-related neglected tropical diseases. Geneva, Switzerland;
446 2022. Contract No.: Licence: CC BY-NC-SA 3.0 IGO.
- 447 9. Yotsu R, Fuller LC, Murdoch ME, Revankar C, Barogui Y, Pemmaraju VRR,
448 et al. WHO strategic framework for integrated control and management of skin-
449 related neglected tropical diseases (skin NTDs). What does this mean for
450 dermatologists? *British Journal of Dermatology*. 2023; 188(2): 157-159.
- 451 10. Yotsu RR, Itoh S, Yao KA, Kouadio K, Ugai K, Koffi YD, et al. The Early
452 Detection and Case Management of Skin Diseases With an mHealth App
453 (eSkinHealth): Protocol for a Mixed Methods Pilot Study in Cote d'Ivoire. *JMIR*
454 *Res Protoc*. 2022;11(9):e39867.

- 455 11. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image
456 recognition. IEEE Conference on Computer Vision and Pattern Recognition;
457 2016; Las Vegas, NV, USA.
- 458 12. Simonyan K, Zisserman A. Very Deep Convolutoinal Networks for Large-
459 scale Image Recognition. arXiv preprint. 2015(1409):1556.
- 460 13. Ding Z, Liu H, editors. Marginalized Latent Sematic Encoder for Zero-Shot
461 Learning. IEEE/CVF Conference on Computer Vision and Pattern Recognition;
462 2019; Long Beach, CA, USA.
- 463 14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L, editors. ImageNet: A
464 large-scale hierarchical image database. IEEE Comput Vis Pattern Recognit;
465 2009; Miami, Florida, USA.
- 466 15. Chicco D, Jurman G. The advantages of the Matthews correlation
467 coefficient (MCC) over F1 score and accuracy in binary classification
468 evaluation. BMC Genomics. 2020;21(1):6.
- 469 16. Abdi H, Williams LJ. Principal component analysis. Wires Computational
470 Statistics. 2010;2(4):433-59.
- 471 17. Barbieri RR, Xu Y, Setian L, Souza-Santos PT, Trivedi A, Cristofono J, et al.
472 Reimagining leprosy elimination with AI analysis of a combination of skin lesion
473 images with demographic and clinical data. Lancet Reg Health Am.
474 2022;9:100192.
- 475 18. Eddyani M, Sopoh GE, Ayelo G, Brun LVC, Roux JJ, Barogui Y, et al.
476 Diagnostic Accuracy of Clinical and Microbiological Signs in Patients With Skin
477 Lesions Resembling Buruli Ulcer in an Endemic Region. Clinical infectious
478 diseases : an official publication of the Infectious Diseases Society of America.
479 2018;67(6):827-34.
- 480 19. van der Werf TS. Diagnostic Tests for Buruli Ulcer: Clinical Judgment
481 Revisited. Clinical infectious diseases : an official publication of the Infectious
482 Diseases Society of America. 2018;67(6):835-6.
- 483 20. Bretzel G, Siegmund V, Nitschke J, Herbinger KH, Thompson W, Klutse E,
484 et al. A stepwise approach to the laboratory diagnosis of Buruli ulcer disease.
485 Trop Med Int Health. 2007;12(1):89-96.
- 486 21. Siegmund V, Adjei O, Nitschke J, Thompson W, Klutse E, Herbinger KH, et
487 al. Dry reagent-based polymerase chain reaction compared with other
488 laboratory methods available for the diagnosis of Buruli ulcer disease. Clinical
489 infectious diseases : an official publication of the Infectious Diseases Society of
490 America. 2007;45(1):68-75.

- 491 22. Jing T, Xu B, Li J, Ding Z. Towards Fair Knowledge Transfer for Imbalanced
492 Domain Adaptation. IEEE Transactions on Image Processing. 2021;30:8200-
493 11.
- 494 23. Hospers IC, Wiersma IC, Dijkstra PU, Stienstra Y, Etuaful S, Ampadu EO,
495 et al. Distribution of Buruli ulcer lesions over body surface area in a large case
496 series in Ghana: uncovering clues for mode of transmission. Trans R Soc Trop
497 Med Hyg. 2005;99(3):196-201.
- 498 24. Sexton-Oates NK, Stewardson AJ, Yerramilli A, Johnson PDR. Does skin
499 surface temperature variation account for Buruli ulcer lesion distribution? PLoS
500 Negl Trop Dis. 2020;14(4):e0007732.
- 501 25. Chang MJ, Lipner SR. Analysis of Skin Color on the American Academy of
502 Dermatology Public Education Website. Journal of drugs in dermatology : JDD.
503 2020;19(12):1236-7.
- 504 26. Okoji UK, Lipoff JB. Demographics of US dermatology residents interested
505 in skin of color: An analysis of website profiles. J Am Acad Dermatol.
506 2021;85(3):786-8.
- 507



medRxiv preprint doi: <https://doi.org/10.1101/2023.03.14.23287243>; this version posted March 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Figure 1. Confusion matrices showing the proportions of diagnosis in the original images [ground truth] and the predicted diagnosis by the deep learning model [prediction]. Left: Setting A, using a model trained with PCR-positive Buruli ulcer cases. Right: Setting B, using a model trained with clinically-diagnosed Buruli ulcer cases.

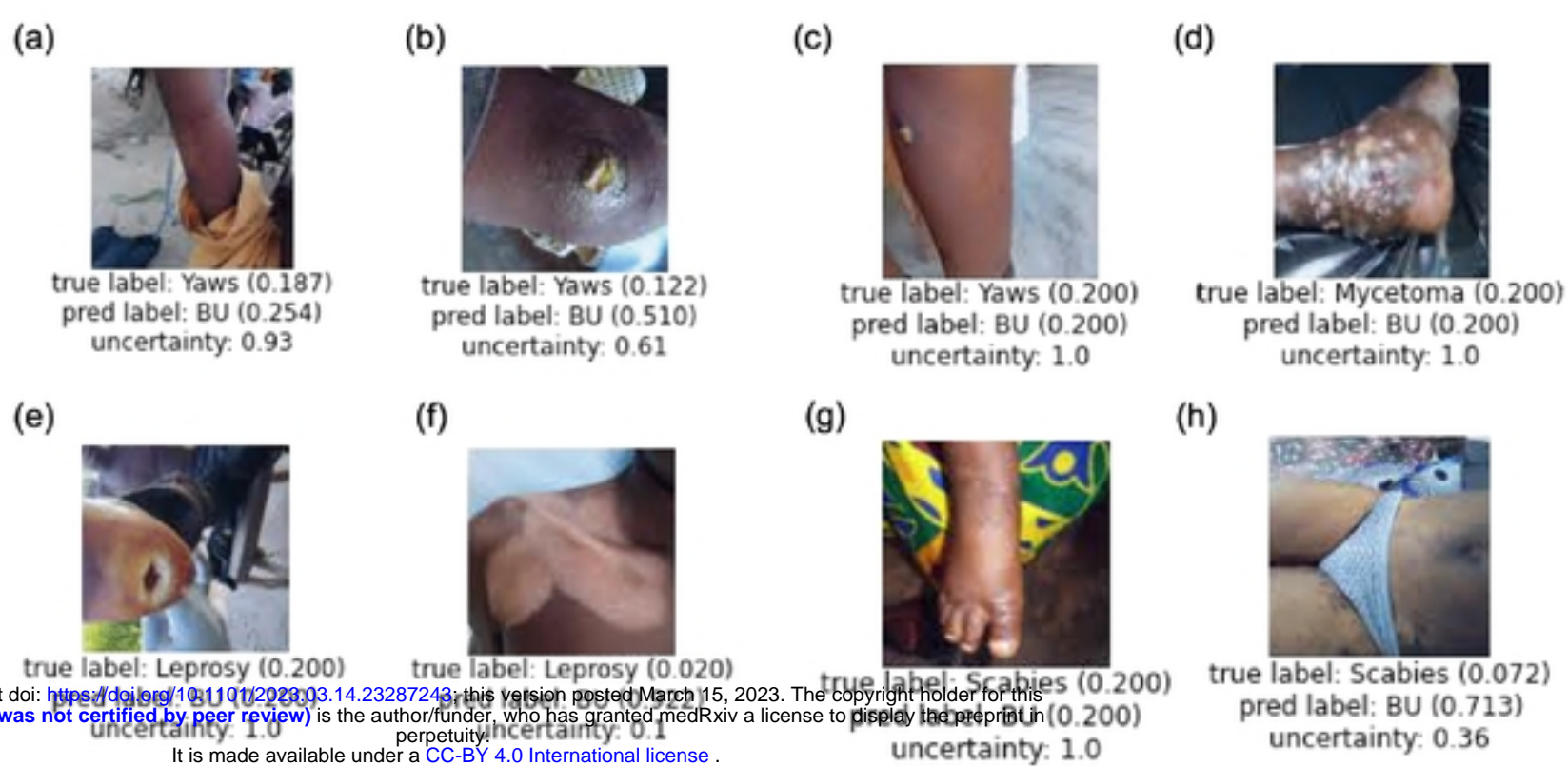
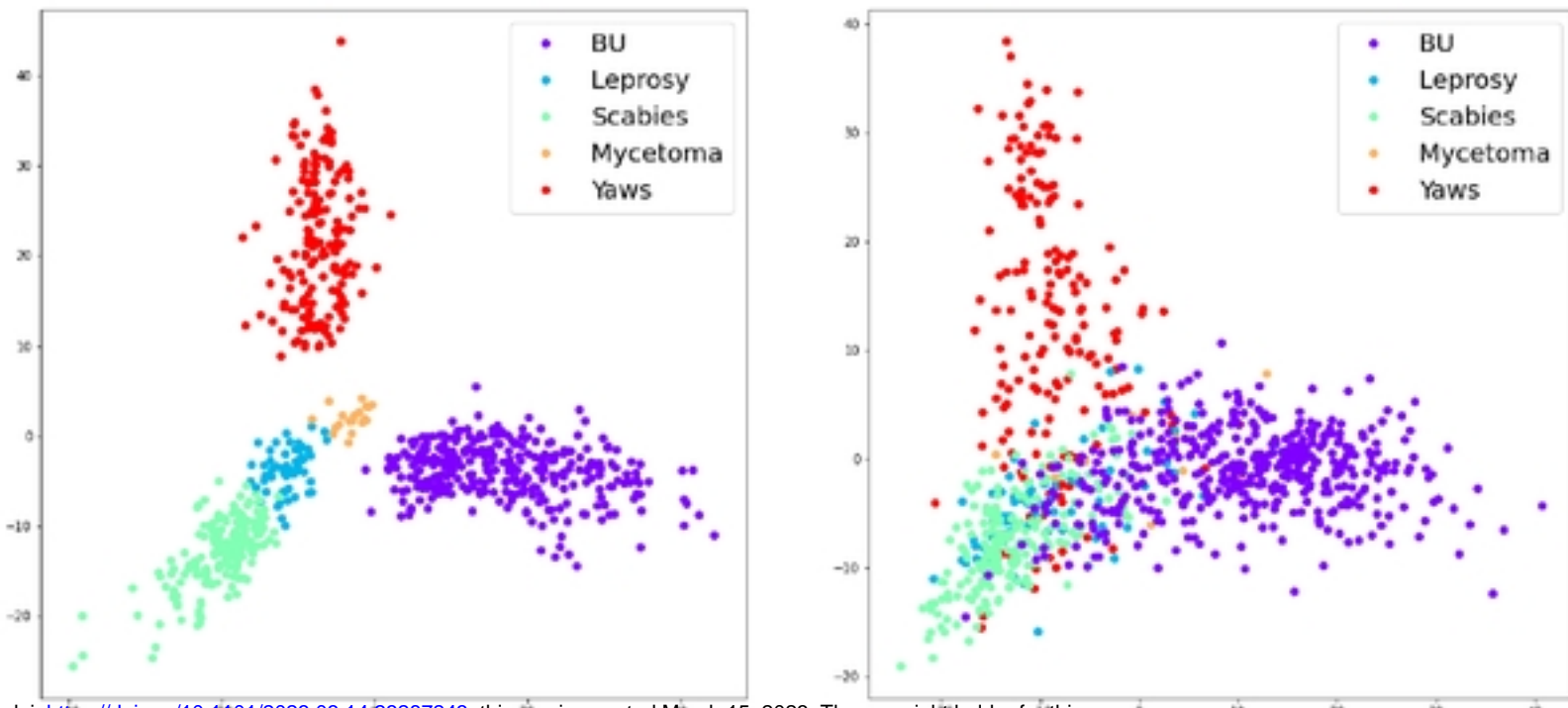


Figure 2. Examples of incorrect prediction made by our AI-based diagnosis model based on ResNet-50, with (k=50)% training data for all data (Task 1).



medRxiv preprint doi: <https://doi.org/10.1101/2023.03.14.23287243>; this version posted March 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Figure 3. Mapping out diagnostic accuracy using Principal Component Analysis.

(a) Training samples from 5 diseases. (b) Test samples from 5 diseases.