# Reimagining leprosy elimination with AI analysis of a combination of skin lesion images with demographic and clinical data

Raquel R Barbieri,[a*,1] Yixi Xu,[b*,1] Lucy Setian,[c] Paulo Thiago Souza-Santos,[a] Anusua Trivedi,[b] Jim Cristofono,[b] Ricardo Bhering,[a] Kevin White,[b] Anna M Sales,[a] Geralyn Miller,[b] José Augusto C Nery,[a] Michael Sharman,[b] Richard Bumann,[b] Shun Zhang,[b] Mohamad Goldust,[d,e] Euzenir N Sarno,[a] Fareed Mirza,[c] Arielle Cavaliero,[c] Sander Timmer,[b] Elena Bonfiglioli,[b] Cairns Smith,[f] David Scollard,[g] Alexander A. Navarini,[d] Ann Aerts,[c] Juan Lavista Ferres,[b*] and Milton O Moraes[a*]

[a]Laboratório de Hanseníase Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Fiocruz, Rio de Janeiro, Brazil
[b]Microsoft, One Microsoft Way, One Microsoft Way, Redmond, WA, United States
[c]Novartis Foundation, Basel, Switzerland
[d]University of Basel, Basel, Switzerland
[e]Department of Dermatology, University Medical Center Mainz, Mainz, Germany
[f]University of Aberdeen, Aberdeen, Scotland
[g]Retired, Wilbraham, MA, United States

The Lancet Regional Health - Americas 2022;9: 100192
Published online 3 February 2022
https://doi.org/10.1016/j.lana.2022.100192

## Summary

**Background** Leprosy is an infectious disease that mostly affects underserved populations. Although it has been largely eliminated, still about 200'000 new patients are diagnosed annually. In the absence of a diagnostic test, clinical diagnosis is often delayed, potentially leading to irreversible neurological damage and its resulting stigma, as well as continued transmission. Accelerating diagnosis could significantly contribute to advancing global leprosy elimination. Digital and Artificial Intelligence (AI) driven technology has shown potential to augment health workers abilities in making faster and more accurate diagnosis, especially when using images such as in the fields of dermatology or ophthalmology. That made us start the quest for an AI-driven diagnosis assistant for leprosy, based on skin images.

**Methods** Here we describe the accuracy of an AI-enabled image-based diagnosis assistant for leprosy, called AI4Leprosy, based on a combination of skin images and clinical data, collected following a standardized process. In a Brazilian leprosy national referral center, 222 patients with leprosy or other dermatological conditions were included, and the 1229 collected skin images and 585 sets of metadata are stored in an open-source dataset for other researchers to exploit.

**Findings** We used this dataset to test whether a CNN-based AI algorithm could contribute to leprosy diagnosis and employed three AI models, testing images and metadata both independently and in combination. AI modeling indicated that the most important clinical signs are thermal sensitivity loss, nodules and papules, feet paresthesia, number of lesions and gender, but also scaling surface and pruritus that were negatively associated with leprosy. Using elastic-net logistic regression provided a high classification accuracy (90%) and an area under curve (AUC) of 96.46% for leprosy diagnosis.

**Interpretation** Future validation of these models is underway, gathering larger datasets from populations of different skin types and collecting images with smartphone cameras to mimic real world settings. We hope that the results of our research will lead to clinical solutions that help accelerate global leprosy elimination.

**Funding** This study was partially funded by Novartis Foundation and Microsoft (in-kind contribution).



**Keywords:** Leprosy; Artificial intelligence; AI; Image-based diagnosis; Dermatology; Skin lesions; AI4leprosy

*Corresponding authors.
*E-mail addresses:* jlavista@microsoft.com (J.L. Ferres), milton.moraes@fiocruz.br (M.O. Moraes).
[1] Co-first authors.

## Research in context

### *Evidence before the study*

According to the WHO Leprosy Guidelines (2018) diagnosis must be based on clinical evaluation, ideally with assistance of slit-skin smears or skin or nerve biopsies histopathological examination. One of the following cardinal signs should be detected such as the presence of the mycobacteria, through acid fast staining testing slit-skin smears, or presence of thickened peripheral nerves. Most of the available laboratory methods exhibit low sensitivity, while qPCR, which exhibits higher sensitivity, still needs reproducibility and standardization with good manufacturing practices. Moreover, all methods from slit-skin smears to biopsies are invasive and laborious requiring infrastructure that is not available in many settings. Early diagnosis is central towards leprosy control, but the delay in diagnosis has been impeding further achievements towards reduction of incidence. Digital health technologies such as AI-powered apps could improve diagnosis of leprosy since it has been successfully applied to the diagnosis of melanoma and other skin conditions.

### *Added value of this study*

From 2018 until 2020, a research protocol for collection of high-resolution images, from leprosy patients and other dermatological diseases along with clinical and demographical data was implemented at FIOCRUZ clinic, Rio de Janeiro, Brazil. Images were stored, labelled and artificial intelligence algorithms were tested. Data show that the AI model's accuracy to detect leprosy previously identified by dermatologists reached 90% and an area under curve (AUC) of 96.46% when elastic-net logistic regression was used. Here we provide a proof-of-concept that AI can be effectively used to determine the probability of leprosy.

### *Implications of all available evidence*

With further independent validation, this AI model may be the first step toward alleviating challenges around skin diseases. This AI model may lead to a non-invasive approach, using a specific set of predictors, that is easy to implement in mobile phones to estimate the probability of leprosy. In the future, this approach may aid health professionals in referring suspect patients to reference centers, when necessary, which in turn provides increased precision to public health strategies on controlling disease burden.

## Introduction

Leprosy is a neglected tropical disease (NTD) for which diagnosis is often delayed because symptoms take between two months to 20 years to appear and disease progression is slow.[1] While histological analysis and qPCR of skin biopsies can aid leprosy diagnosis, as does Ziehl-Neelsen staining for detection of *Mycobacterium leprae* on slit skin smears from ear lobes,[2] there is currently no diagnostic test considered the gold standard for leprosy. Up to date, leprosy diagnosis remains a clinical one.[1] Waning leprosy expertise amongst health workers, and delayed healthcare seeking caused by leprosy-associated stigma, often results in delayed diagnosis, when neurological damage has already taken place, leading to irreversible disabilities.

Although the widespread availability of free multidrug therapy (MDT) secured a 99% reduction of the global leprosy burden, there are still up to 200'000 patients newly diagnosed annually, and new tools are needed to cover the last miles towards leprosy elimination,[3] defined as zero transmission. One recently validated strategy is contact tracing followed by post-exposure-prophylaxis (PEP) for contact persons of leprosy patients, recommended by WHO following the successful implementation of the Leprosy PEP initiative.[4]

Clinical presentation of leprosy depends on a patient's specific immunity. The World Health Organization (WHO) classifies the disease into paucibacillary (PB) and multibacillary (MB) forms, for people presenting with 1-5 or more than 6 skin lesions, respectively.[5] Compared to a standard classification for research purposes,[6] the PB leprosy group includes tuberculoid (TT) and borderline tuberculoid (BT) patients, while the MB group includes patients with borderline-borderline (BB), borderline-lepromatous (BL), and lepromatous (LL) leprosy. Differential diagnosis for PB leprosy includes dermatological manifestations, such as those caused by pityriasis alba, syphilis, psoriasis, granuloma annulare, and sarcoidosis, while MB leprosy and specifically the LL forms need to be differentiated from diseases such as lymphoma, neurofibromatosis and xanthoma.

As previously demonstrated, digital technology can accelerate leprosy diagnosis, allowing frontline health workers to send mobile phone skin images to a reference dermatologist for guidance and referral.[7] Increasing evidence around machine learning enabling faster and more accurate image-based diagnosis in disciplines such as radiology, pathology, and dermatology, motivated us to develop an Artificial Intelligence (AI) driven "diagnosis accelerator" for leprosy, AI4Leprosy, using a combination of skin images, clinical and reported symptoms.

Most of the AI-driven diagnosis evidence in dermatology comes from melanoma, and algorithms such as deep convolutional neural networks (CNN) have delivered comparable accuracy to dermatologists, in differentiating malignant from benign lesions.[8] Deep neural networks have shown to exceed specialists in differentiating melanoma from benign mimickers such as nevi and seborrheic keratoses on dermoscopy images.[9] Larger training datasets and advances in algorithm

development further increased its performance,[10] while augmenting human performance for lesions where physicians reported low diagnostic confidence.[11]

This paper describes the process of developing an AI model that analyzes skin lesions, metadata and de-identified patient information to determine the probability of leprosy, explaining the acquisition and compilation of images in a high-resolution image open-source database, their CNN analysis and combination with clinical data. Further, we describe the methodology for storage of skin images and their labeling as leprosy or leprosy-like lesions. Our de-identified images include a diversity of clinical manifestations and skin types and are openly available in a dataverse repository for further use by researchers and healthcare workers. Combining images, clinical and demographic data from confirmed leprosy patients and patients with leprosy-like skin lesions, we trained AI algorithms to differentiate leprosy from other conditions. To our knowledge, this is the first large open-source image- and databank available for leprosy, and AI model for suspecting leprosy. Our results indicate that AI modeling can work as a powerful tool to accelerate and increase accuracy of leprosy suspection for further diagnostics confirmation.

## Methods

### Ethics

Data and image collection were conducted at the leprosy clinic of the Oswaldo Cruz Foundation in Rio de Janeiro, a national reference center receiving a diversity of patients and their families from all over Brazil. Patient interviews, clinical consultations and image collections followed standard operational processes. The study was approved by the Ethics and Research Committee of the Oswaldo Cruz Foundation CAAE: 38053314.2.0000.5248, number: 976.330–10/03/2015. All study participants provided informed consent and could request clarification about the research or its risks and benefits at any time. Images from skin lesions were de-identified and follow up photos could be taken to evaluate disease improvement or progression.

### Study design and workflow

Patient inclusion and exclusion criteria and overall workflow until AI data analysis, are illustrated in Supplementary Fig. 1. At the first visit, patients presenting skin lesions were examined and those diagnosed with "leprosy-like lesions" (macule, plaques or nodules) were eligible to enroll. These patients received information on the research, completed the informed consent as well as general health and demographic information along with potential symptoms. Photography was taken and then skin biopsies and slit skin smears from ear lobes were used to confirm (or exclude) leprosy diagnosis. Leprosy diagnosis was defined only after the results of histology and Ziehl-Neelsen staining of slit skin smears for the detection of leprosy tissue morphological features or *Mycobacterium leprae,* respectively. qPCR was only performed when equivocal results in histological analysis were found.[2] Final diagnosis was only established at the third visit (up to 30 days after visit two), as laboratory results would then be available and enable the dermatologists to communicate the diagnosis and initiate the appropriate treatment. Leprosy diagnosis followed the WHO operational classification and the Ridley and Jopling[6] (Supplementary Figs. 2 and 3). Confirmed leprosy patients were treated with MDT, according to the national guidelines of the Ministry of Health and WHO.[5]

Patients' socio-demographic data were recorded in the standard local health information system, including a tag (name and register number) to identify study patients. Skin lesion images were stored in a computer exclusively used for the study and on a backup device, and labeled with the VGG image annotator.[12] Image assessments followed minimal requirements of the International Skin Imaging Collaboration (ISIC) including background color, lighting, field of view, focus/depth of field, resolution, scale, and color calibration. Details on the step-by-step process for image capture, and storage are described in Supplementary Fig. 4. Overall, up to three images were taken from each skin lesion: a panoramic photo to identify the body part where the lesion was situated, a close-up photo and another picture from the edge of the lesion, including surrounding normal skin. The DOI for repository can be accessed at: https://doi.org/10.35078/1PSIEL. The code can also be seen here: https://github.com/microsoft/leprosy-skin-lesion-ai-analysis.

To keep the original features, images were stored in raw format and .jpeg extensions were generated for transfer onto Microsoft Azure™. Specific study metadata describing demographics, clinical information and skin lesions details were entered into Epi info™ [13] and images were uploaded on the cloud to develop the AI algorithms.

### Inclusion and exclusion criteria

All patients with leprosy-like skin lesions and above the age of six were included, except those who did not provide informed consent, or who were diagnosed with HIV or tuberculosis. Parental consent was required for minors under 18 years old and patients who did not present for their final diagnosis at visit 3, were excluded.

### Clinical examination and data collection

The dermatologists assessed inclusion and exclusion criteria before inviting patients to participate in the

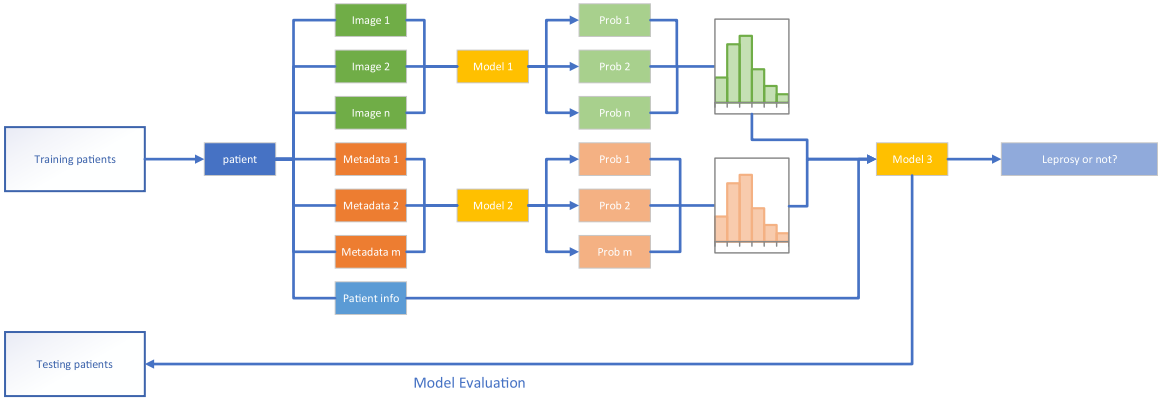


**Figure 1.** Data modelling overview. Of the 228 recruited patients, 222 patients were finally included in the data analysis. Images (model 1) or metadata (model 2) from 182 patients were used to train the algorithms in a training dataset, while 40 patients were separated as an independent testing group only used for validation in model 3. The outputs of models 1 and 2 were histograms that fed the accuracy and area under curve (AUC) calculations.

study, completed the study forms and linked each skin lesion image to the corresponding patient metadata.

During the consultation, dermatologists assessed symptoms at the lesion sites, such as pain, itching, sensitivity loss or hyperesthesia, and checked for the presence of paresthesia in the hands and feet. Peripheral nerve thickening was also evaluated by dermato-neurological examination. In addition, dermatologists assessed whether there was diffuse skin infiltration, loss of eyebrows, lichenification or scaling at the lesion surface. Color (erythematous, hypo-or hyperpigmented) and body location (upper limbs, lower limbs, or trunk) of skin lesions were recorded, as well as their type (macule, papule, plaque, nodule, ulcer, vesicle, or blister). The diameter of skin lesions was measured and for lesions $\geq$ 4 cm, the temperature was assessed on the lesion and on the adjacent healthy skin area, as well as contralaterally. Lastly, sensitivity (thermal, pain and tactile) was checked for those skin lesions that were larger than 1.5 cm, resulting in nodules e.g., not being evaluated for temperature or sensitivity. Sensitivity in leprosy skin lesions was assessed using a standard clinical test performed during the dermatological examination. To evaluate thermal sensitivity tubes containing warm and cold water are used, a sharp stick or needle was used to assess the sensation of pain, and the tactile sensation was assessed using a piece of cotton. Demographic and descriptive tables were plotted for all variables, using continuous or discrete statistics to compare demographic and metadata variables.

## Data modeling

As the predictive models we used require equal input dimensions, we developed a two-step patient-level model, first predicting the probability of leprosy based on the skin lesion image (Model 1) or the metadata (Model 2). Each model produced a probability of leprosy for each image or set of metadata, as shown in Figure 1. Given that patients could have multiple lesions or metadata records, we combined outputs from both models per patient in a histogram, to represent the predicted probabilities. Lastly, Model 3 was trained to combine analysis made in the first step, with the patient information. This last step established the overall probability by combining the histograms from Model 1 and 2, with patient information.

## Preprocessing data

When any of the variables had too few observations, they were grouped with the most frequently occurring variable. For example, we combine "Nodule" and "Papule" into a new group for the variable "type". If the lesions were too small, the diameter and temperatures were not measured and the diameter would be imputed as 1 cm, while the temperature would be replaced by the median of the patient's temperatures. If no temperature could be measured for a patient, the median of all patients' temperatures would be used. The value "missing" became an additional category for sensory loss of lesions that were too small to measure sensory function. Predictors that mostly consisted of a single value were dropped.

## Role of funding source

The Novartis Foundation established the concept of AI4leprosy, provided technical input into the design phase of the initiative, and ensured overall coordination of the initiative, while it was not involved in the data analysis. Microsoft provided data science expertise and in-kind cloud credit for the development of the model.

## Results

A total of 228 patients diagnosed with leprosy or other dermatological manifestations were recruited. Three patients, who did not adhere to the consultations schedule and failed to confirm their diagnosis (visit 3), were excluded. Age, gender, and general clinical features of the 225 included patients are presented in Supplementary Table 1. Because of the absence of close-up images, three patients were excluded, resulting in a final 222 study patients (Supplementary Fig. 5). From those, a total of 582 skin lesions and 1226 images were collected, uploaded, and stored on the cloud. Loss of eyebrows or eyelashes, diffuse skin infiltration, enlarged nerves and foot paresthesia were more common in patients with confirmed leprosy, while lichenification and scaling surface of skin lesions were more often seen in patients with other dermatologic conditions (Supplementary Table 1). All the models were evaluated by comparing the estimated probability of leprosy with the previous diagnoses made by dermatologists.

### Learning algorithms

We first split the 222 patients into two groups. One group of randomly selected 40 patients, was used as the testing patients to validate the final patient-level models. The remaining 182 patients were used for both model training and selection. For each experiment of Model 1 or 2, we used 5-fold cross-validation to evaluate the performance of each algorithm (Supplementary Fig. 6).

For Model 1, we combined the following settings (*i*) the neural network architecture: Inception-v4 or ResNet-50; (*ii*) tuning strategy: tune all (fine-tune the complete neural net model) or freeze (train only the output layer); (*iii*) input image type: close-up only or all images; (*iv*) optimizer: stochastic gradient descent. A fine-tuned ResNet-50 using close-up images gave the best cross-validated accuracy (ACC) as well as the area under curve, AUC (Table 1). This was the final algorithm to train Model 1, which generated the probability histogram to feed Model 3, as shown in Figure 1. The performance of the ResNet-50 was superior (66.6% accuracy and 74.56% for AUC), although lower accuracy and AUC was achieved when both close-up and other images were included (Table 1).

For Model 2 we tested the three machine learning methods, elastic-net logistic regression (LR), XGBoost (XGB), random forests (RF) (Table 2). Temperatures and diameters are time-consuming to measure in practice. Thus, we evaluated the impact of removing these two features by creating a subset of features without temperatures and diameters. The data demonstrated that elastic-net logistic regression using the subset features achieved the highest AUC score (88.6%). This was the final algorithm to train Model 2, which generated the probability histogram to feed Model 3.

Removing temperature and diameter of skin lesions only minimally influenced performance of the algorithms. To interpret the model, we used elastic-net logistic regression with repeated 10-fold cross validation, on the complete dataset including the subset features. The model selects the following variables: type, color, site, sensory loss, thermal, tactile, pruritus, hyperesthesia, and asymptomatic as shown

| Mean ACC & AUC | | ResNet-50 | | Inception-v4 | |
|---|---|---|---|---|---|
| | | Tune All | Freeze | Tune All | Freeze |
| All | ACC (SD)* | 0.6138 (0.040) | 0.5723 (0.051) | 0.5828 (0.070) | 0.5212 (0.050) |
| | AUC (SD) | 0.6760 (0.057) | 0.6003 (0.094) | 0.6144 (0.106) | 0.5487 (0.045) |
| Close-up | ACC (SD) | 0.6660 (0.099) | 0.5790 (0.064) | 0.5834 (0.092) | 0.5661 (0.059) |
| | AUC (SD) | 0.7456 (0.113) | 0.6542 (0.104) | 0.6590 (0.099) | 0.5919 (0.089) |

***Table 1*: The performance comparison for Model 1 (images only) using ResNEt-50 and Inception-v4 neural network architectures.**
* ACC – accuracy; AUC- area under curve; SD – standard deviation.

| Mean score | Elastic-net Regression (LR) | | XGBoost (XGB) | | Random forest (RF) | |
|---|---|---|---|---|---|---|
| | Full | Subset | Full | Subset | Full | Subset |
| ACC (SD) | 0.813 (0.058) | 0.817 (0.06) | 0.808 (0.086) | 0.818 (0.075) | 0.779 (0.088) | 0.818 (0.073) |
| AUC (SD) | 0.881 (0.082) | 0.880 (0.080) | 0.849 (0.086) | 0.846 (0.092) | 0.836 (0.071) | 0.863 (0.090) |
| Sensitivity (SD) | 0.841 (0.118) | 0.845 (0.115) | 0.818 (0.067) | 0.85 (0.092) | 0.795 (0.129) | 0.845 (0.090) |
| Specificity (SD) | 0.791 (0.173) | 0.794 (0.174) | 0.79 (0.177) | 0.784 (0.155) | 0.763 (0.177) | 0.789 (0.158) |

***Table 2*: The performance comparison for Model 2 using features extracted from the form of the lesion for Elastic-net logistic regression (LR), XGBoost (XGB) and Random forests (RF) machine learning methods. For the full data analysis, data included 15 predictors collected at the clinical evaluation as described in the methods section. For subset analysis, temperature and diameter features were excluded.**

| | Elastic-net Regression-LR (%) | | | | XGBoost - XGB (%) | | | | Random Forest - RF (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC (SE) | AUC (SE) | SEN (SE) | SP (SE) | ACC (SE) | AUC (SE) | SEN (SE) | SP (SE) | ACC (SE) | AUC (SE) | SEN (SE) | SP (SE) |
| Model 1 outputs | 72 (0.070) | 73 (0.083) | 78 (0.098) | 68 (0.099) | 65 (0.075) | 71 (0.086) | 67 (0.111) | 64 (0.102) | 65 (0.075) | 71 (0.084) | 67 (0.111) | 64 (0.102) |
| Model 2 outputs | 92 (0.041) | 95 (0.025) | 94 (0.054) | 91 (0.061) | 90 (0.047) | 92 (0.03) | 89 (0.074) | 91 (0.061) | 90 (0.047) | 92 (0.039) | 89 (0.074) | 91 (0.061) |
| Patient info | 88 (0.052) | 96 (0.020) | 72 (0.105) | 100 (0) | 95 (0.034) | 99 (0.006) | 94 (0.054) | 95 (0.044) | 95 (0.034) | 98 (0.01) | 89 (0.074) | 1 (0) |
| Model 1 & 2 outputs | 80 (0.063) | 89 (0.054) | 83 (0.087) | 77 (0.089) | 80 (0.063) | 88 (0.053) | 78 (0.098) | 82 (0.082) | 82 (0.060) | 89 (0.054) | 78 (0.09) | 86 (0.073) |
| Model 1 outputs + patient info | 78 (0.066) | 86 (0.061) | 78 (0.09) | 77 (0.089) | 65 (0.075) | 79 (0.071) | 72 (0.106) | 59 (0.104) | 75 (0.068) | 85 (0.065) | 72 (0.105) | 77 (0.089) |
| Model 2 outputs + patient info | 90 (0.047) | 96 (0.019) | 89 (0.074) | 91 (0.061) | 88 (0.052) | 93 (0.034) | 83 (0.088) | 91 (0.061) | 88 (0.052) | 96 (0.021) | 83 (0.087) | 91 (0.061) |
| All | 88 (0.052) | 92 (0.039) | 83 (0.087) | 91 (0.061) | 78 (0.066) | 87 (0.058) | 72 (0.106) | 82 (0.082) | 80 (0.063) | 90 (0.049) | 72 (0.105) | 86 (0.073) |

***Table 3*: The performance comparison for Model 3 using the all the training patients and logistic regression (LR), XGBoost (XGB) and random forest (RF). All metrics were obtained by validating the models on a separate testing dataset from 40 patients. Results are shown for Models 1 and 2 separately or combined. Patient info means the information collected on the patient information document.**

*ACC-accuracy; AUC area under curve; SEN-sensitivity; SP-specificity; SE-standard error.

in Table 3, which OR estimates are summarized in Supplementary Table 2.

The findings of this model are aligned with clinical observations, given that e.g., sensory loss is a typical feature of leprosy or that leprosy lesions rarely cause pruritus.

For Model 3, the other two models were retrained with the complete set of 182 patients, using ResNet-50, respectively elastic-net logistic regression algorithms. These two were the best algorithms selected by 5-fold cross-validation in the first step. Then we trained Model 3 using all the 182 patients by elastic-net logistic regression, XGBoost and Random Forest. The final patient-level models were validated on the 40-testing patients. Random Forest using patient information alone delivered the highest AUC – 98.74% (Table 3). For XGBoost and Random Forest, there is little benefit of including Model 1 and Model 2 outputs, in addition to the patient information. However, for elastic-net logistic regression, the inclusion of Model 2 outputs increases the AUC as compared to the model built on patient information alone (Table 3).

XGBoost and Random Forest are better at capturing nonlinear relationships than elastic-net logistic regression, and this could be the reason why the inclusion of Model 2 outputs increases the AUC for elastic-net logistic regression but not for XGBoost or Random Forest. In addition, higher input dimension could lead to overestimation of the algorithm accuracy, especially for complex machine learning models. Such overfitted model tend to produce over-optimistic results, which overstate its predictive power. Finally, overlapping fields existed between the metadata and patient information such as sensory loss or pruritus. Although values were not the same, they characterized similar aspects: for example, 'nodule' would be recorded both on the lesion document and the patient information. As for the models built on a single data resource, using Model 1 outputs alone yields a moderate AUC around 70%, whereas using Model 2 outputs or patient information gives a strong model with an AUC > 90%.

As shown in Table 2, Random Forest using patient information alone achieves the highest AUC on the testing patients. The ten most important features we identified were: thermal sensitivity loss, nodules and papules, feet paresthesia, number of lesions, gender, scaling surface, pruritus, trunk, no symptoms in the skin lesion and diffuse infiltration (Figure 2). The permutation feature importance was measured by the decrease in model accuracy when changing the feature's values. Interestingly, the odds ratio (OR) estimates using model 2 by elastic-net logistic regression with clinical information confirmed some of these variables exhibiting the highest OR values (Supplementary Table 2).

Because of the small sample size, we chose to also use elastic-net logistic regression with Model 2 outputs and patient information. This offered a high

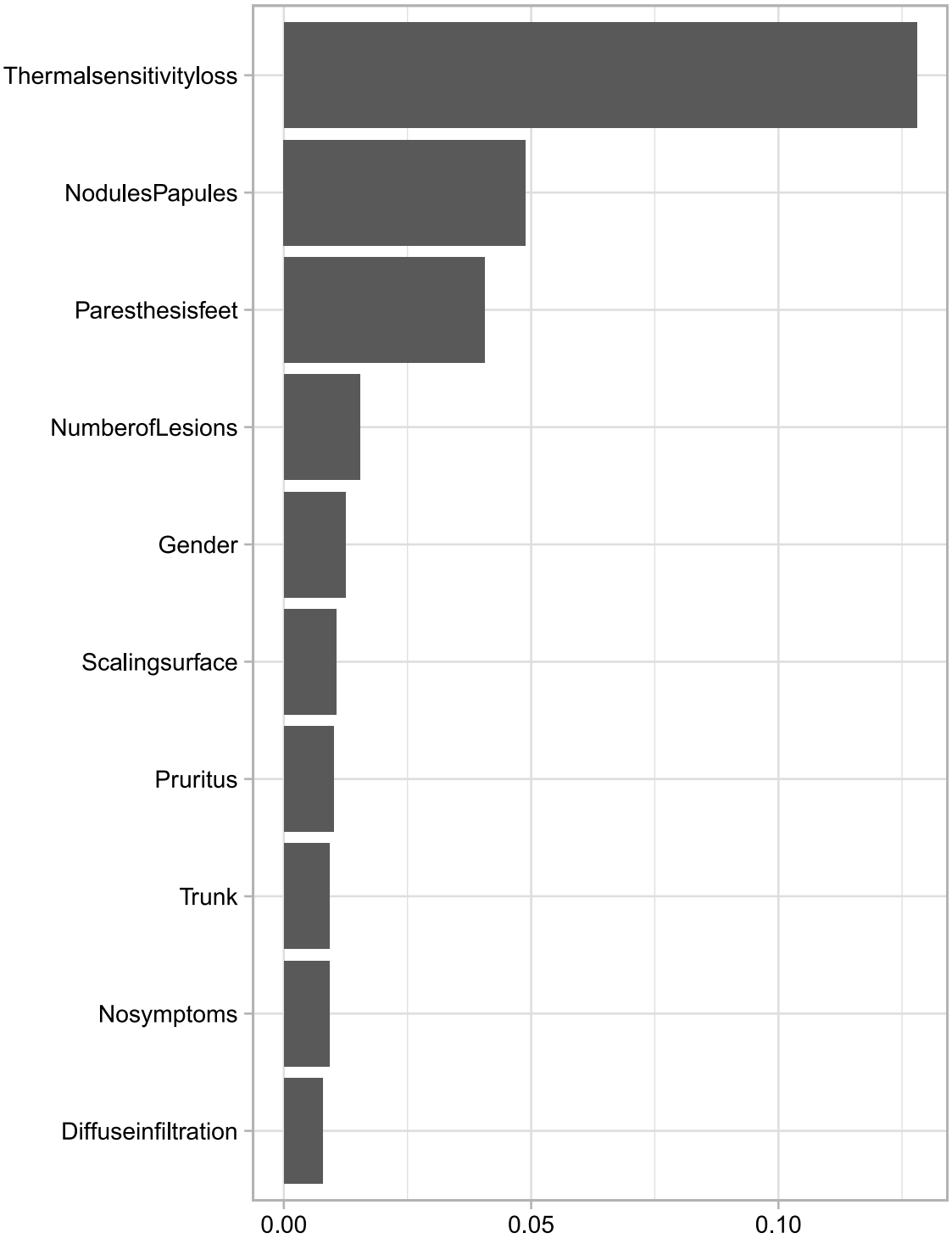


**Figure 2.** Permutation feature value, representing the most important features after model 3 training.

classification accuracy (90%) and AUC (96.46%) and was simpler and more interpretable than XGBoost or Random Forest. To define the final model, we used elastic-net logistic regression with repeated 10-fold cross validation on the complete dataset including the testing patients which is summarized in Supplementary Table 3.

The top 10 features described in Figure 2 were selected by logistic regression except nodules, papules, pruritus, and the absence of symptoms. The three remaining features were selected by Model 2 and the probability histogram was used by logistic regression. Thus, both random forests and elastic-net logistic regression have a similar preference with regards to the variables they use.

## Discussion

Our study was designed to provide an open-source dataset of high-quality skin images and clinical data to evaluate the feasibility and accuracy of AI driven leprosy analysis model based on skin images and clinical symptoms. We are not aware of other such existing datasets and hope that our work can be an inspiration for further training of AI models in dermatology. As several other NTDs present skin manifestations, the model described here could be an example for replication and assist other research on NTDs, which can be debilitating diseases and often affect the poorest populations in the world. We understand that data security and privacy is of major concern, and, therefore, we defined a protocol avoiding individual recognition. Although we provide access to the dataset, it will be restricted and made available only upon registration and user validation.

The results of our research indicate that probability models were able to recognize leprosy with high accuracy (96.4%), especially when combining Elastic-net Regression model 2 outputs and patient information. Another CNN-based AI dermatology diagnosis assistant (AIDDA) delivers 89.4% accuracy for the diagnosis of psoriasis and 92.57% for atopic dermatitis and eczema. AIDDA's availability on a smartphone app has proven to significantly increase diagnostic accuracy for less specialized health professionals.[14] Based on evidence from AIDDA research, in the next phase of research on AI4Leprosy, we will train the algorithm by collecting images and data through a smartphone app, to help validate and further improve the model on lower quality images and ultimately better mimic real-world settings. This rollout is foreseen to happen in remote rural and urban environments in Asian and African settings, with the aim to remediate any selection bias that would have been introduced in our study by collecting data in a leprosy reference center. Training the model in frontline settings and on other skin types will be essential, even though Brazil has a quite diverse population that allowed image collection from a variety of skin types.

A previous study from India also tested skin image-based leprosy diagnosis with AI,[15] however, we were unable to trace the origin of the images and verify the dataset labels. As it is extremely common for leprosy patients to present multiple skin lesions, ignoring that fact could result in the model trying to memorize patterns of skin and providing overoptimistic results prediction. A rigorous image collection and processing protocol is a prerequisite to help prevent the introduction of systematic biases in the algorithms.[8]

Not surprisingly, a cardinal sign for leprosy such as sensitivity loss contributed significantly to the AI algorithm. This symptom plus the features of diffuse skin infiltration, enlarged nerves, hand or foot paresthesia and hyperesthesia, and the loss of eyebrows or eyelashes, were ideal to build a logistic model that was aligned with clinical observations and provide an accurate estimate. Besides those, lichenification and scaling surfaces were also incorporated in the model, as signs that are mainly indicative of non-leprosy lesions. The presence of skin lesions in more than one site significantly increased the predicted probability for leprosy. On the other hand, our study demonstrates that the model experiences more difficulties in detecting leprosy from skin images than from the metadata or patient information. Type, colour, and sensitivity of the lesion, for example, are all descriptions introduced by human experts, who know that these features are useful in the diagnosis of leprosy. If we would feed the model exclusively with images and not with clinical descriptions, the algorithm would have to learn this by itself. Contrarily to our observations, others have demonstrated rather good correlations between dermatologic data and visual characteristics of the skin lesions.[16] Although we understand that collecting clinical data can be challenging and would need training, we developed an application to help data collectors adequately use the algorithm, further increasing its applicability. This application is not meant to be a diagnostic tool, but one that helps the clinician or healthcare worker to identify lesions suspect of leprosy and as such can accelerate referral to health professionals for proper diagnosis.

A model will only work well when additional data from new patients from multiple geographical backgrounds are used to continue to train and improve it. This is possible when further data collection follows a similar protocol to that used for the model's training, enhanced by digital data collection tools to prevent manual errors. Model transferability in different populations must be considered.[17] In 2018, Han et al.[18] trained their skin lesion model on an Asian population and tested it on a European population. This resulted in an accuracy of 55.7% over the 10-classes of Dermfit,[18] which was significantly lower than that of other models trained and tested over Dermfit (81%).[19] Such a decrease in the accuracy can be attributed to the differences in skin manifestations across populations or the lack of

transferability in learned features across datasets, due to image acquisition protocols. For that reason, it will be paramount to train the AI4leprosy model in multiple, diverse populations.

### Contributors

The concept of AI4Leprosy was established by AA, MOM, JLF, RRB, AMS, AT, ENS, AC, ST, FM, EB, and GM and its full development was coordinated by LS, JLF, GM, RB, EB and JC. Data and image collection was performed by RRB, RB, PTSS, AMS, JACN, ES. Experiments were run by RB1, ST, RRB, PTSS, YX, RB2 and MS, while data curation was performed by RRB, PTSS and YX; data analysis and modelling was done by YX, MOM, JLF, RRB, PTSS and MS. Results analysis and global protocol design were performed by LS, MG, AAN, JLF with contributions from KW, JC, AT, CS, DS, AC, AA. Leprosy expertise was provided by CS, DS, AC and AA. The full draft was developed by RRB, YX, MOM, JLF and AA. All authors reviewed and approved the final draft.

## Funding

This study was funded by the Novartis Foundation, with in-kind expertise from Microsoft AI4Health.

### Declaration of interests

We declare no competing interests.

### Acknowledgements

We express our gratitude to John Kahan from Microsoft for championing and driving this partnership as from the start and for his unrivalled support to the entire initiative. We thank Johannes Boch from the Novartis Foundation for supporting the coordination of this initiative at several points in time. And we are grateful to all patients for their willingness to participate in the study.

### Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.lana.2022.100192.

#### References

1 Cooreman E, Gillini L, Pemmaraju V, et al. Guidelines for the diagnosis, treatment and prevention of leprosy. *World Heal Organ.* 2018;1.

2 Barbieri RR, Manta FSN, Moreira SJM, et al. Quantitative polymerase chain reaction in paucibacillary leprosy diagnosis: a follow-up study. *PLoS Negl Trop Dis.* 2019;13(3). https://doi.org/10.1371/journal.pntd.0007147.

3 OMS. Weekly epidemiological record. Global leprosy update, 2018: moving towards a leprosy. *Wkly Epidemiol Rec.* 2019;94.

4 Richardus JH, Tiwari A, Barth-Jaeggi T, et al. Leprosy post-exposure prophylaxis with single-dose rifampicin (LPEP): an international feasibility programme. *Lancet Glob Heal.* 2021;9(1):e81–e90. https://doi.org/10.1016/S2214-109X(20)30396-X.

5 *WHO Expert Committee on Leprosy (1997 : Geneva, Switzerland) & World Health Organization. WHO Expert Committee on Leprosy : Seventh Report.* World Health Organization; 1998. https://apps.who.int/iris/handle/10665/42060.

6 Ridley DS, Jopling WH. Classification of leprosy according to immunity. A five-group system. *Int J Lepr Other Mycobact Dis.* 1966;34(3).

7 Mieras LF, Taal AT, Post EB, Ndeve AGZ, van Hees CLM. The development of a mobile application to support peripheral health workers to diagnose and treat people with skin diseases in resource-poor settings. *Trop Med Infect Dis.* 2018;3(3). https://doi.org/10.3390/tropicalmed3030102.

8 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–118. https://doi.org/10.1038/nature21056.

9 Brinker TJ, Hekler A, Enk AH, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur J Cancer.* 2019;111. https://doi.org/10.1016/j.ejca.2019.02.005.

10 Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the international skin imaging collaboration 2017. *J Am Acad Dermatol.* 2020;82(3). https://doi.org/10.1016/j.jaad.2019.07.016.

11 Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol.* 2018;78. https://doi.org/10.1016/j.jaad.2017.08.016.

12 Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: *Proceedings of the MM 2019 - 27th ACM International Conference on Multimedia.* 2019. https://doi.org/10.1145/3343031.3350535.

13 Dean AG, Arner TG, Sunki G, et al. Epi InfoTM, a database and statistics program for public health professionals. 2011.

14 Wu H, Yin H, Chen H, et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Ann Transl Med.* 2020;8(9). https://doi.org/10.21037/atm.2020.04.39.

15 Baweja HS, Parhar T. Leprosy lesion recognition using convolutional neural networks. In: *Proceedings of the International Conference on Machine Learning and Cybernetics.* 1, 2016. https://doi.org/10.1109/ICMLC.2016.7860891.

16 Yang J, Sun X, Liang J, Rosin PL. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2018. https://doi.org/10.1109/CVPR.2018.00137.

17 Kawahara J, Hamarneh G. Visual diagnosis of dermatological disorders: human and machine performance. *arXiv.* Published online 2019.

18 Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* 2018;138(7). https://doi.org/10.1016/j.jid.2018.01.028.

19 Kawahara J, Bentaieb A, Hamarneh G. Deep features to classify skin lesions. In: *Proceedings of the International Symposium on Biomedical Imaging.* 2016. https://doi.org/10.1109/ISBI.2016.7493528.