A Web-scrapped Skin Image Database of Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles

Towhidul Islam[1†], Mohammad Arafat Hussain[2*†], Forhad Uddin Hasan Chowdhury[3] and B. M. Riazul Islam[4]

[1]Department of Computer Science and Engineering, Northern University Bangladesh, Dhaka, 1215, Bangladesh.
[2*]Boston Children's Hospital, Harvard Medical School, Boston, 02115, Massachusetts, USA.
[3]Department of Medicine, Dhaka Medical College Hospital, Dhaka, 1000, Bangladesh.
[4]Health Information Unit, Directorate General of Health Services, Dhaka, 1212, Bangladesh.

*Corresponding author(s). E-mail(s): mohammad.hussain@childrens.harvard.edu;
Contributing authors: towhidul.islam@nub.ac.bd; drmarufsomc@gmail.com; riaz.somc@mis.dghs.gov.bd;
†These authors contributed equally to this work.

Abstract

Monkeypox has emerged as a fast-spreading disease around the world and an outbreak has been reported in 42 countries so far. Although the clinical attributes of Monkeypox are similar to that of Smallpox, skin lesions and rashes caused by Monkeypox often resemble that of other pox types, e.g., Chickenpox and Cowpox. This scenario makes an early diagnosis of Monkeypox challenging for the healthcare professional just by observing the visual appearance of lesions and rashes. The rarity of Monkeypox before the current outbreak further created a knowledge gap among healthcare professionals around the world. To tackle this challenging situation, scientists are taking motivation from the success of supervised machine learning in COVID-19 detection. However, the lack of Monkeypox skin image data is making the bottleneck of using machine learning in Monkeypox detection from skin images of patients. Therefore, in this project, we introduce the Monkeypox Skin Image Dataset (MSID), the largest of its kind so far. We used web-scrapping to collect Monkeypox, Chickenpox, Smallpox, Cowpox and Measles infected skin as well as healthy skin images to build a comprehensive image database and made it publicly available. We believe that our database will facilitate the development of baseline machine learning algorithms for early Monkeypox detection in clinical settings. Our dataset is available at the following Kaggle link: https://www.kaggle.com/datasets/arafathussain/monkeypox-skin-image-dataset-2022.

Keywords: Monkeypox, Chickenpox, Smallpox, Cowpox, Measles, machine learning, deep learning, image data, skin lesions, CNN

1 Introduction

While the world is gradually recovering from the havoc caused by the Coronavirus disease (COVID-19), another infectious disease, known as Monkeypox, has been spreading around the world at a fast pace. To date, an outbreak of Monkeypox has been reported in 42 countries.[1] Human infection of Monkeypox virus was first reported in the Democratic Republic of Congo (formerly Zaire) in 1970 (Thornhill et al, 2022), which was transmitted to humans from animals. Monkeypox virus belongs to the genus *Orthopoxvirus* of the family *Poxviridae* (Shchelkunov et al, 2002), which shows similar symptoms to that of Smallpox (Thornhill et al, 2022). Smallpox has been eradicated in 1970, which led to the cessation of Smallpox vaccination. Since then, Monkeypox is considered the most important *Orthopoxvirus* for human health. Monkeypox used to be primarily reported in the African continent, however, it has been widely spreading in urban areas in different locations of the world (Thornhill et al, 2022). The current outbreak of Monkeypox in humans on a global scale is believed to be due to the changes in Monkeypox's biological attributes, changes in human behavior, or both.[2]

The clinical attributes of Monkeypox are similar to that of Smallpox.[3]. In addition, Monkeypox skin lesions and rashes often resemble that of other pox types, e.g., Chickenpox and Cowpox, making an early diagnosis of Monkeypox challenging for the healthcare professional. The rarity of Monkeypox before the current outbreak (Sklenovska and Van Ranst, 2018) also created a knowledge gap among healthcare professionals around the world. Polymerase chain reaction (PCR) is typically considered the most accurate tool for the Monkeypox detection (Erez et al, 2019), however, healthcare professionals are accustomed to diagnosing pox infections by visual observation of skin rash

[1] https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON393

[2] https://www.cdc.gov/poxvirus/monkeypox/response/2022/world-map.html

[3] https://www.who.int/news-room/fact-sheets/detail/monkeypox

and lesion. Despite a low mortality rate from Monkeypox infection (i.e., 1%-10%) (Gong et al, 2022), early detection of Monkeypox can facilitate isolation of patients as well as contact tracing for effective containment of a community spread of Monkeypox.

Different machine learning (ML), especially deep learning (DL), approaches played a significant role in COVID-19 detection and severity analysis from images of different medical imaging modalities (e.g., computed tomography (CT), chest X-ray, and chest ultrasound) (Sun et al, 2022; Akbarimajd et al, 2022; Momeny et al, 2021). This success motivates the use of ML or DL approaches to detect Monkeypox from the skin images of patients. However, supervised or semi-supervised learning approaches are data-driven and require a large number of data to well-train an ML or DL model. Unfortunately, there is no publicly available healthcare facility or infectious disease control authority-disclosed database of Monkeypox skin lesions or rash. In such a situation, web-scrapping (i.e., extracting data from websites) (Dogucu and Çetinkaya-Rundel, 2021) to collect Monkeypox skin lesion images may be the only alternative to facilitate the development of ML- and DL-based Monkeypox infection detection algorithms.

In this project, we used web-scrapping to collect Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles infected skin as well as healthy skin images to build a comprehensive image database and made it publicly available. To date, to our knowledge, there are two other web-scrapping-based Monkeypox databases currently available (Ahsan et al, 2022; Ali et al, 2022). However, the number of data and the number of pox classes are limited in those databases. Our database is unique in the following way:

1. Our database contains skin lesion/rash images of five different diseases, i.e., Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles, as well as contains healthy skin images.
2. This database contains more web-scrapped pox and measles image data, before augmentation, compared to other similar databases.
3. We enhanced the privacy of patients in images by blocking exposed eyes and private parts with black boxes.
4. We used various augmentation techniques to increase the number of data by 49-times.
5. We acknowledged the sources of all images in our database and presented the source/credit list as supplementary material.

2 Methodology

In this section, we briefly describe our approach to data collection, expert screening, and data augmentation. We also show the pipeline of our database development in Fig. 1.

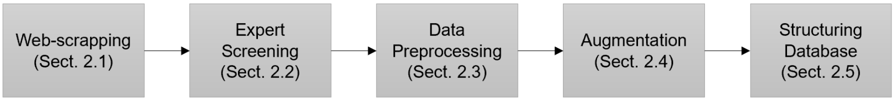

Web-scrapping (Sect. 2.1)
Expert Screening (Sect. 2.2)
Data Preprocessing (Sect. 2.3)
Augmentation (Sect. 2.4)
Structuring Database (Sect. 2.5)

Fig. 1: The pipeline of our database development. Each block mentions the corresponding section number in this paper, where the details are presented.

2.1 Web-scrapping for Data Collection

As we mentioned in section 1 that a publicly available healthcare facility or infectious disease control authority-disclosed database of Monkeypox skin lesions or rashes is yet to come, we used web-scrapping to collect Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles infected skin as well as healthy skin images. We use the Google search engine to search for different types of pox-infected skin images, Measles-infected skin images, and healthy skin images from various sources such as websites, news portals, blogs, and image portals. We modified our search for collecting those images that fall under "Creative Commons licenses." However, for several pox cases, we hardly find images under "Creative Commons licenses," thus collected images that fall under "Commercial & other licenses." That is why, we include a list as supplementary material that includes the uniform resource locator (URL) of source, access date, and photo credit (if any) for all our collected images. In Fig. 2, we show some example images from our database.

2.2 Expert Screening of Data

After collecting image data, two expert physicians, who specialized in infectious diseases, sequentially screened all the images to validate the supposed infection. In Fig. 3, we show a pie chart of the percentage of original web-scrapped image data per class in our database, after the expert screening.

2.3 Data Preprocessing

In this step, we cropped images manually to remove unwanted peripheral background regions. Further, to make patients non-identifiable from their corresponding images, we blocked the eye region with black boxes. We also did the same to hide revealed private parts as much as possible, without blocking the skin lesions/rashes. Since typical DL algorithms expect an input image of square shape in terms of pixel counts (often 224×224×3 pixels), we added extra blank pixels in the periphery of many images to avoid excessive stretching of the actual skin lesions during image resizing. After that, we cropped and resized all the images to 224×224×3 pixels (3-channel RGB format).

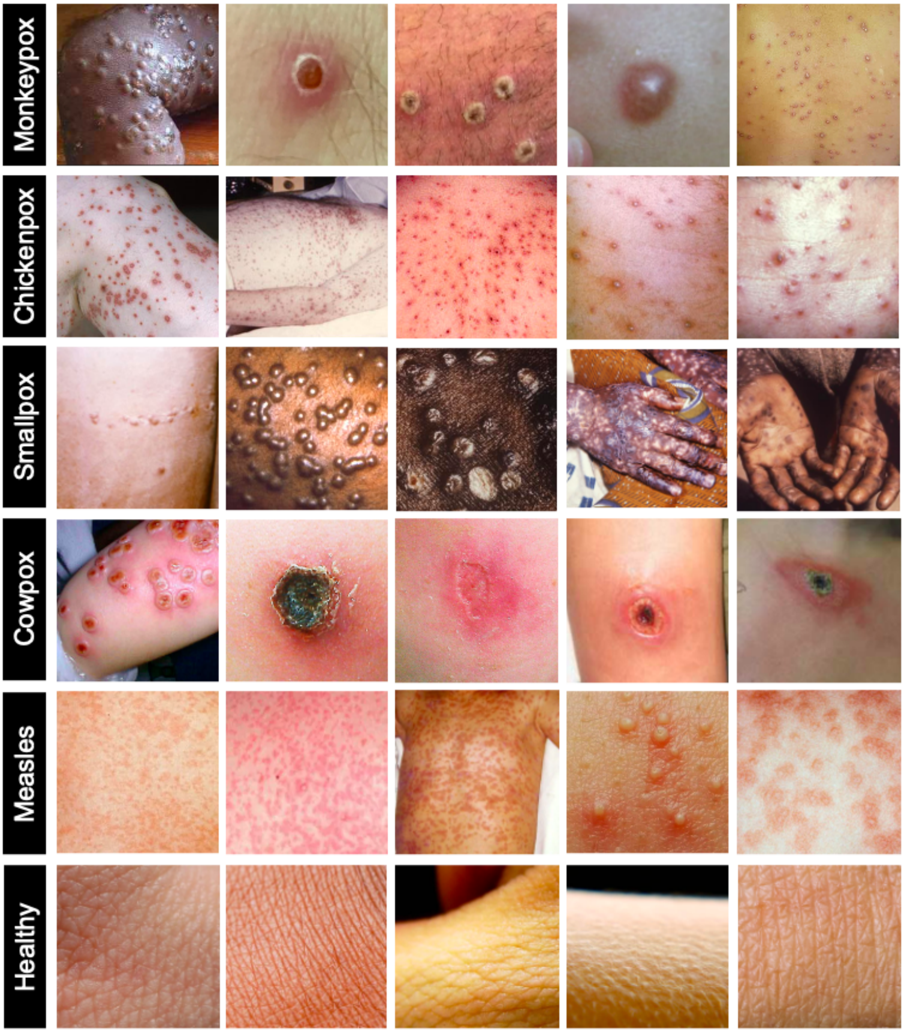


Fig. 2: Example skin images of Monkeypox, Chickenpox, Smallpox, Cowpox, Measles, and healthy cases (first to sixth rows, respectively) from our database.

2.4 Augmentation

To increase the number of images as well as introduce variability in data, we performed the following 19 augmentation operations on the web-scrapped image data using Python imaging library (PIL) version 9.2.0, and scikit-image library version 0.19.3:

1. Brightness modification with a randomly generated factor (range [0.5, 2]).
2. Color modification with a randomly generated factor (range [0.5, 1.5]).
3. Sharpness modification with a randomly generated factor (range [0.5, 2]).

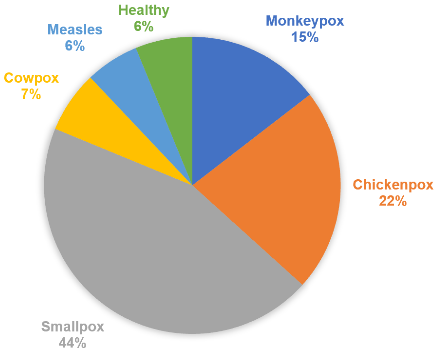


Fig. 3: A pie chart showing the percentage of original web-scrapped image data per class in our database.

4. Image translation along height and width with a randomly generated distance between -25 and 25 pixels.
5. Image shearing along height and width with randomly generated parameters.
6. Adding Gaussian noise of zero mean and randomly generated variance (range [0.005, 0.02]) to images.
7. Adding Speckle noise of zero mean and randomly generated variance (range [0.005, 0.02]) to images.
8. Adding salt noise to randomly generated number of image pixels (range [2%, 8%]).
9. Adding pepper noise to randomly generated number of image pixels (range [2%, 8%]).
10. Adding salt & pepper noise to randomly generated number of image pixels (range [2%, 8%]).
11. Modifying image pixels values based on local variance.
12. Blurring an image by a Gaussian kernel with randomly generated radius (range [1, 3] pixels).
13. Contrast modification with a randomly generated factor (range [1, 1.5]).
14. Rotating all images by 90°.
15. Rotating images by a random angle (range [-45°, 45°]).
16. Zooming in an image by about 9%.
17. Zooming in an image by about 18%.
18. Flipping images along the height.
19. Flipping images along the width.

In Table 1, we show the number of original and augmented images per class in our database. We also show a flow diagram of our augmentation pipeline in Fig. 4. We used these augmentation operations to increase the data by 49×.

Table 1: Distribution of image classes in the Monkeypox Skin Image Dataset.

Class label	**No. of Original Images**	**No. of Augmented Images**
Monkeypox	117	5,733
Chickenpox	178	8,722
Smallpox	358	17,542
Cowpox	54	2,646
Measles	47	2,303
Healthy	50	2,450
Total	804	39,396

2.5 Naming Convention and Database Structure

We named the preprocessed original images as "xx_yyyy.jpg," where "xx" is either of "mo," "ch," "sm," "co," "me," or "he," representing Monkeypox, Chickenpox, Smallpox, Cowpox, Measles, or healthy, respectively, and "yyyy" denotes the serial of an image belonging to "xx" class. After augmentation, images in the pool are named "aug_xx_yyyy_zzzz.jpg," where "zzzz" represents the serial of the augmented image of original image "yyyy" belonging to "xx" class. Note that "aug_xx_yyyy_0001.jpg" is always the preprocessed but resized original image. So, during validation and testing/inference, it will be sufficient to call "aug_xx_yyyy_0001.jpg" for accuracy estimation. Also, during data partitioning for training, validation, and testing, or cross-validation, all "zzzz" of a particular "yyyy" should be in a specific fold (i.e., either training or validation or testing) to avoid cross-fold data splitting of the same patient "yyyy".

We structured our database in a way that makes it easy to call in an ML or DL model training routine. In Fig. 5, we show the structure of our database in terms of directory map.

3 Expected Outcomes

We expect that our proposed database will help to achieve the following objectives:

Baseline ML and DL Algorithms: We expect that our database will facilitate the development of baseline ML and DL algorithms for early Monkeypox detection until a larger dataset from any national or international government or autonomous entity/authority comes out.

Achieving Better Detection-ability: Unlike other similar databases, our database contains image data for more pox types as well as Measles and healthy skins. In addition, it contains more images per class than that in other databases. Therefore, we expect that training any ML or DL model on our dataset would achieve a better ability to classify different types of skin lesions/rashes.

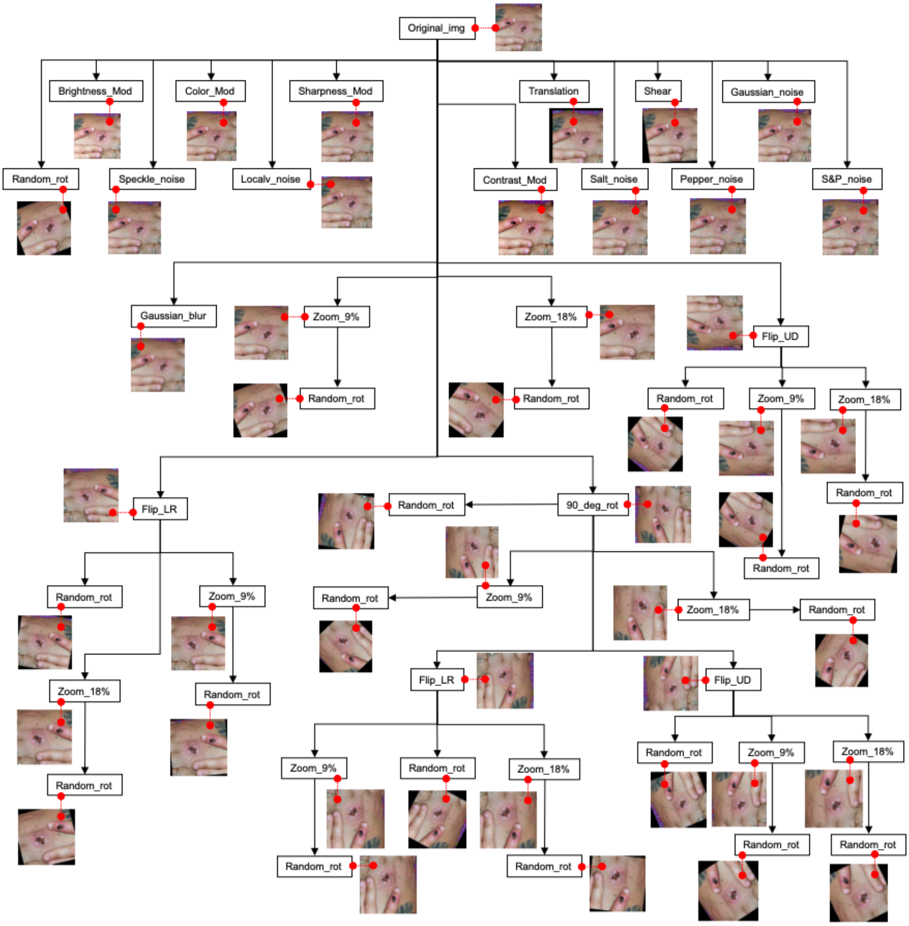


Fig. 4: Flow diagram of our augmentation pipeline. Using this pipeline of operations, we increased the data by 49×. We also show representative augmented images attached to each box. Acronyms- Original_img: original image, Brightness_Mod: brightness modification, Color_Mod: color modification, Sharpness_Mod: sharpness modification, Gaussian_noise: Gaussian noise addition, Random_rot: random rotation, Speckle_noise: Speckle noise addition, Localv_noise: local variance noise addition, Contrast_Mod: contrast modification, Salt_noise: salt noise addition, Pepper_noise: pepper noise addition, S&P_noise: salt & pepper noise addition, Gaussian_blur: Gaussian blurring, Zoom_9%: 9% zooming in, Zoom_18%: 18% zooming in, Flip_LR: flipping along the width, Flip_UD: flipping along the height, and 90_deg_rot: rotation of image by 90°.

Avoiding Model Over-fitting: Since our database has multi-class images, which were again multiplied 49× by augmentation, we expect that any ML

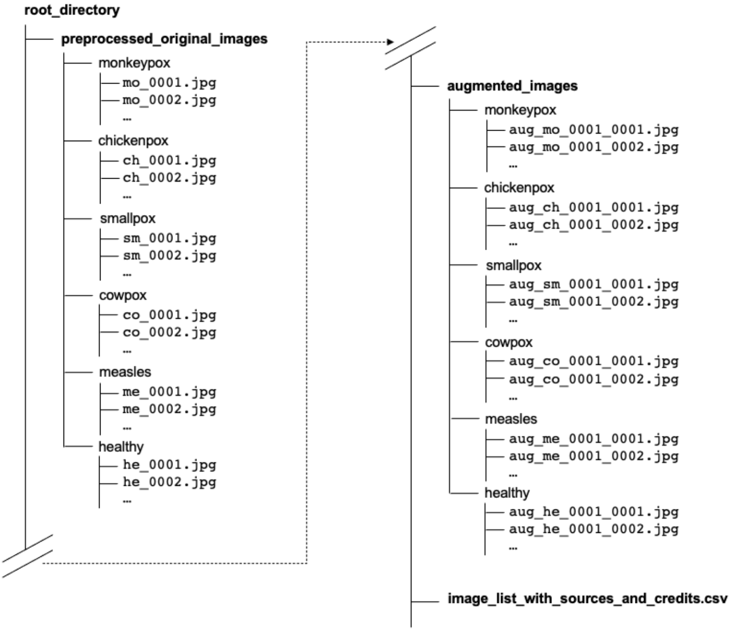


Fig. 5: Directory map of our database.

or DL model would be better generalized (i.e., not over-fitted on a specific disease class) if trained on our dataset.

Aid in Clinical Environment: An ML or DL model, trained on our dataset, can help in the clinical diagnosis of Monkeypox. It may reduce the dependency on conventional microscopic image analysis and PCR-based Monkeypox detection. It may also reduce the close contact between healthcare professionals and patients. In fact, a patient can be diagnosed by a healthcare professional remotely by analyzing a skin lesion image of the patient taken by a smartphone.

4 Conclusion

In this project, we used web-scrapping to build a comprehensive database of Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles infected skin images. Compared to other similar databases, ours has the largest number of actual images per class as well as the largest number of augmented images per class. We took expert opinions from two physicians to validate the disease in an image. We also made a list to acknowledge the source of each image and made it available for public use and scrutiny. We believe that this database will facilitate the development of baseline ML and DL algorithms for early Monkeypox detection, empower ML or DL models to achieve a better ability as well as generalizability to classify different types of skin lesions/rashes, and overall help in the clinical diagnosis of Monkeypox. We also aim to keep this database updated with new data as soon as they appear on the web.

Supplementary Material
The list of the image sources for all the classes can be found in this link: https://www.kaggle.com/datasets/arafathussain/monkeypox-skin-image-dataset-2022?select=image_list_with_sources_and_credits.xlsx

References

Ahsan MM, Uddin MR, Luna SA (2022) Monkeypox image data collection. arXiv preprint arXiv:220601774

Akbarimajd A, Hoertel N, Hussain MA, et al (2022) Learning-to-augment incorporated noise-robust deep cnn for detection of covid-19 in noisy x-ray images. Journal of Computational Science p 101763

Ali SN, Ahmed M, Paul J, et al (2022) Monkeypox skin lesion detection using deep learning models: A feasibility study. arXiv preprint arXiv:220703342

Dogucu M, Çetinkaya-Rundel M (2021) Web scraping in the statistics and data science curriculum: Challenges and opportunities. Journal of Statistics and Data Science Education 29(sup1):S112–S122

Erez N, Achdout H, Milrot E, et al (2019) Diagnosis of imported monkeypox, israel, 2018. Emerging infectious diseases 25(5):980

Gong Q, Wang C, Chuai X, et al (2022) Monkeypox virus: a re-emergent threat to humans. Virologica Sinica

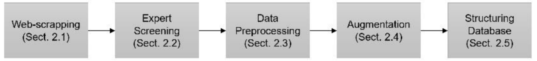
Momeny M, Neshat AA, Hussain MA, et al (2021) Learning-to-augment strategy using noisy and denoised data: Improving generalizability of deep cnn for the detection of covid-19 in x-ray images. Computers in Biology and Medicine 136:104,704

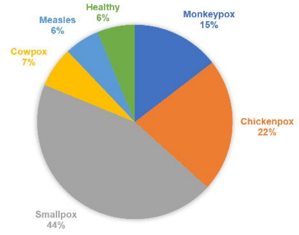
Shchelkunov S, Totmenin A, Safronov P, et al (2002) Analysis of the monkeypox virus genome. Virology 297(2):172–194

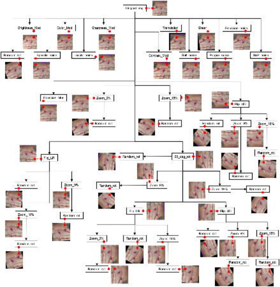
Sklenovska N, Van Ranst M (2018) Emergence of monkeypox as the most important orthopoxvirus infection in humans. Frontiers in public health 6:241

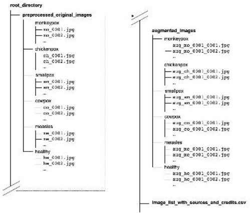
Sun J, Peng L, Li T, et al (2022) Performance of a chest radiograph ai diagnostic tool for covid-19: A prospective observational study. Radiology: Artificial Intelligence 4(4):e210,217

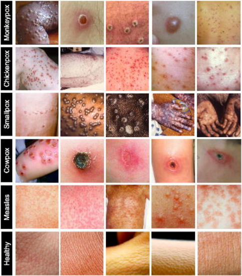
Thornhill JP, Barkati S, Walmsley S, et al (2022) Monkeypox virus infection in humans across 16 countries—april–june 2022. New England Journal of Medicine


Web-scrapping (Sect. 2.1)
Expert Screening (Sect. 2.2)
Data Preprocessing (Sect. 2.3)
Augmentation (Sect. 2.4)
Structuring Database (Sect. 2.5)


Healthy
6%
Monkeypox
15%
Measles
6%
Cowpox
7%
Chickenpox
22%
Smallpox
44%


Original_img
Brightness_Mod
Color_Mod
Sharpness_Mod
Translation
Shear
Gaussian_noise
Random_rot
Speckle_noise
Local_noise
Contrast_Mod
Salt_noise
Pepper_noise
S&P_noise
Gaussian_blur
Zoom_2%
Random_rot
Zoom_15%
Random_rot
Flip_UD
Random_rot
Zoom_4%
Zoom_16%
Random_rot
Random_rot
Flip_LR
Random_rot
30_deg_rot
Random_rot
Zoom_9%
Zoom_16%
Random_rot
Random_rot
Zoom_4%
Random_rot
Zoom_16%
Random_rot
Flip_LR
Flip_UD
Zoom_2%
Random_rot
Zoom_16%
Random_rot
Zoom_4%
Zoom_15%
Random_rot
Random_rot
Random_rot
Random_rot


root_directory
preprocessed_original_images
monkeypox
mo_0001.jpg
mo_0002.jpg
...
chickenpox
ch_0001.jpg
ch_0002.jpg
...
smallpox
sm_0001.jpg
sm_0002.jpg
...
cowpox
co_0001.jpg
co_0002.jpg
...
measles
me_0001.jpg
me_0002.jpg
...
healthy
he_0001.jpg
he_0002.jpg
...
augmented_images
monkeypox
aug_mo_0001_0001.jpg
aug_mo_0001_0002.jpg
...
chickenpox
aug_ch_0001_0001.jpg
aug_ch_0001_0002.jpg
...
smallpox
aug_sm_0001_0001.jpg
aug_sm_0001_0002.jpg
...
cowpox
aug_co_0001_0001.jpg
aug_co_0001_0002.jpg
...
measles
aug_me_0001_0001.jpg
aug_me_0001_0002.jpg
...
healthy
aug_he_0001_0001.jpg
aug_he_0001_0002.jpg
...
image_list_with_sources_and_credits.csv


Monkeypox
Chickenpox
Smallpox
Cowpox
Measles
Healthy