Data in Brief 57 (2024) 110997

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

E-Staining DermaRepo: H&E whole slide image staining dataset

Muhammad Zeeshan Asaf[a], Anum Abdul Salam[a,*], Samavia Khan[b,c], Noah Musolff[c], Muhammad Usman Akram[a], Babar Rao[b,c]

[a] Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[b] Centre for Dermatology, Rutgers Robert Wood Johnson Medical School, Somerset, NJ 08873, USA
[c] Rao Dermatology, 900 Broadway, New York, NY 10003, USA

ARTICLE INFO

Article history:
Received 25 August 2024
Revised 23 September 2024
Accepted 27 September 2024
Available online 5 October 2024

Dataset link: E-Staining DermaRepo (Original data)

Keywords:
Whole slide image segmentation
Bright field microscope
Histological staining
Virtual staining

ABSTRACT

In the era of artificial intelligence and machine learning, computer-aided diagnostic frameworks are data-hungry and require large amounts of annotated data to automate the disease diagnosis procedure. Moreover, to enhance the performance and accuracy of disease diagnosis, procedures need to be automated to ensure timely and accurate diagnosis. We are providing a whole slide image repository comprising unstained skin biopsy images acquired using a bright-field microscope, along with Hematoxylin and Eosin chemically and virtually stained image samples, to virtualize the staining procedure and enhance the efficiency of the disease diagnosis pipeline. The dataset was utilized to train a Dual Contrastive GAN to generate virtually stained image samples. The trained model achieved an FID score of 80.47 between virtually stained and chemically stained image samples, indicating a high correlation of content between synthesized and original images. In contrast, FID scores of 342.01 and 320.40 were observed between unstained images and virtually stained slides, and between unstained images and

* Corresponding author.
E-mail address: anum.abdulsalam@ceme.nust.edu.pk (A.A. Salam).

https://doi.org/10.1016/j.dib.2024.110997

chemically stained images, respectively, indicating less similarity in content.

Specifications Table

Subject	Artificial Intelligence; Computational Pathology; Computer Science; Biomedical Imaging
Specific subject area	Virtual staining of whole slide images acquired using a brightfield microscope enhances structural patterns and disease-associated biomarkers, reducing the time required for disease diagnosis.
Data format	Original whole slide H&E-stained images in .TIFF paired with their associated virtually H&E-stained and chemically H&E-stained images in .PNG
Type of data	Images (Stained and unstained)
Data collection	The data samples were collected by Non-Invasive Diagnostic Innovations (NIDI) in Skin through biopsies, followed by application of H&E chemical staining to highlight disease-associated biomarkers and various skin structural features beneficial for skin disease diagnosis. The stained image samples were then captured using a standard brightfield microscope at a magnification rate of 10x. The corresponding virtually H&E-stained image samples were generated using Dual Contrastive GANs.
Parameters for data collection	The acquired whole slide image samples represent skin biopsies from 15 males and 7 females, with an age range of 34–83 years and a median age of 67.71 years. Images were de-identified to protect patient privacy.
Description for data collection	Unstained image samples paired with chemically stained image samples were provided by Non-Invasive Diagnostics Innovations (NIDI) in Skin. Images were captured by conducting a skin biopsy, followed by chemical staining of the extracted sample. The resulting stained and unstained samples were captured using a standard brightfield microscope at 10x magnification and high resolution, producing images ranging from 0.22 Gigapixels to approximately 1.7 Gigapixels.
Data source location	**Institution:** BioMedical Image and Signal Analysis (BIOMISA) research group, Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Non-Invasive Diagnostic Innovations (NIDI) **City:** Rawalpindi and Islamabad; New York **Country:** Pakistan; USA
Data accessibility	**Repository name:** Data identification number: 10.17632/gxgg933ny3.1 Direct URL to data: https://data.mendeley.com/datasets/gxgg933ny3/1 Instructions for accessing these data samples: The data repository is public and can be accessed by anyone for research purpose. Please cite this paper whenever used in a research work. Utilisation of data for any commercial purpose is restricted.
Related research article	Asaf, Muhammad Zeeshan, Babar Rao, Muhammad Usman Akram, Sajid Gul Khawaja, Samavia Khan, Thu Minh Truong, Palveen Sekhon, Irfan J. Khan, and Muhammad Shahmir Abbasi. "Dual contrastive learning based image-to-image translation of unstained skin tissue into virtually stained H&E images." *Scientific Reports* 14, no. 1 (2024): 2335.

1. Value of the Data

- High-resolution whole slide images, along with their associated chemically and virtually stained image samples, hold crucial importance in computer-aided diagnosis. These images will aid in model training to replace the tedious chemical staining process with virtual staining, thereby enhancing the efficacy of the disease diagnosis pipeline.
- The collected data samples will aid the researchers in comparing and benchmarking their proposed frameworks [1,2] in the field of computational pathology. This will enable researchers to standardize results and ensure reproducibility.
- The dataset has the potential to incorporate interdisciplinary research across pathology, computer science, and bioinformatics, promoting collaboration and innovation in the development of new diagnostic tools and techniques.

- The data repository will promote the use of virtual staining in disease diagnosis, thereby minimizing the need for chemical reagents and manual labor. Virtual staining has the potential to enhance lab efficiency, enabling labs to manage resources more effectively and process higher case volumes.

2. Background

Dermatological conditions impose a significant burden within the global health landscape. In the Global Burden of Disease (GBD) Studies conducted in 2010 and 2013, they collectively constituted the world's fourth leading cause of non-fatal burden [3,4], and in the most recent study, they were ranked as the eighth leading cause [5]. These conditions impact individuals across all age demographics and represent one of the most prevalent motivations for individuals to seek assistance from healthcare professionals. Dermatological diseases affect individuals of all age groups and skin tones worldwide [6], regardless of their financial status. Disparities in access to healthcare services significantly influence the early detection, accurate diagnosis, and overall treatment outcomes. To control the widespread of skin diseases, there is a need to improve the diagnosis pipeline. Whole slide Imaging is being widely used in skin disease diagnosis and analysis, however when acquired the images needs staining which may take up to 24 h to get the image ready for analysis by pathologist. Moreover, the process is highly dependent on an individual's skills and might differ from an individual to another effecting the process of diagnosis. To address this margin, skin disease diagnosis pipeline can be improved by replacing the process of chemical staining with virtual staining. We aim to provide a dataset to aid in model training, where our dataset is comprised of un-stained and stained pairs.

3. Data Description

The data repository comprises 87 H&E-stained whole slide images (.JPG) located in the "Unstained" directory, along with their associated chemically stained images (.JPG) in the "C_Stained" directory and virtually stained images (.JPG) in the "V_Stained" directory. All directories are organized within the root directory named "H&E-Staining dataset". The images in each category - unstained, chemically stained and virtually stained - are identically named to ensure correspondence between them. For example, the first image in all three directories is titled "HC21-01338(A3-1).10X", indicating their association and making it easier to retrieve the unstained image along with its associated chemically and virtually stained counterparts. Fig. 1 illustrates a whole slide unstained image along with its associated chemically and virtually stained samples.

4. Experimental Design, Materials and Methods

Skin conditions are estimated to affect around 1.8 billion people globally at any given time. In tropical and resource-limited settings, conditions caused by bacterial, viral, fungal, or parasitic infections, are the leading contributors to illness. According to the Global Burden of Disease (GBD) studies from 2010 to 2013, skin diseases were the fourth leading cause of non-fatal burden, while a 2021 study ranked them eighth among all causes [7]. Furthermore, the progression of skin diseases imposes a significant economic burden globally. To mitigate this economic impact and halt disease progression, there is a need for autonomous disease diagnosis frameworks, which necessitate extensive datasets for model training and evaluation.

NIDIskin, America provided unstained and chemically H&E-stained whole slide image samples acquired from skin biopsies using a standard brightfield microscope at 10x magnification of 15 males and 7 females, aged 34–83 years, with an average range of 67.71 years. The images are of high resolution with pixel values between approximately 0.22–1.7 Gigapixels. Data

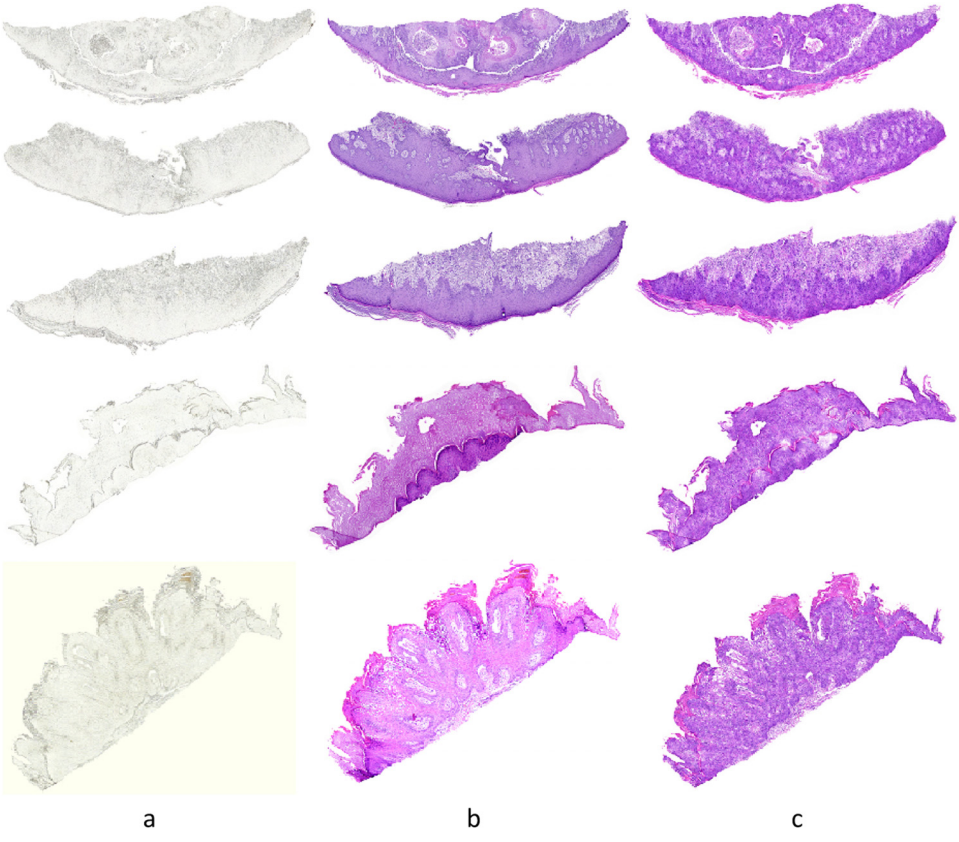

Fig. 1. Data samples from the acquired dataset; (a) Unstained whole slide image acquired using a brightfield microscope; (b) Chemically stained whole slide image, stained using Haematoxylin and Eosin chemicals and acquired using a brightfield microscope; (c) Virtually stained image generated using dual contrastive GANs trained on the unstained and chemically stained image pairs.

Table 1
Dataset description and details.

Attribute	Value
Data Acquisition	Brightfield Microscope
Gender Distribution	15 males, 7 females
Median age	67.71 years
Resolution	0.22 Gigapixels to approx. 1.7 Gigapixels
Image samples	87

acquisition details and descriptions are provided in Table 1. We utilized this dataset to generate virtually H&E-stained image samples through a dual contrastive generative adversarial network [8].

The model was trained on pre-processed stained and unstained patches using a combination loss function that included cross entropy loss L_{ce}, Identity loss L_{id} and GAN loss L_{gan} as illustrated in Fig. 2. The pair of trained models generates H&E-stained images from unstained images and vice versa. The Fréchet Inception Distance (FID) scores for these images are 80.47 for virtually stained versus chemically stained images, 320.4 for chemically stained versus unstained images, and 342.01 for virtually stained versus unstained images, as detailed in Table 2. These results

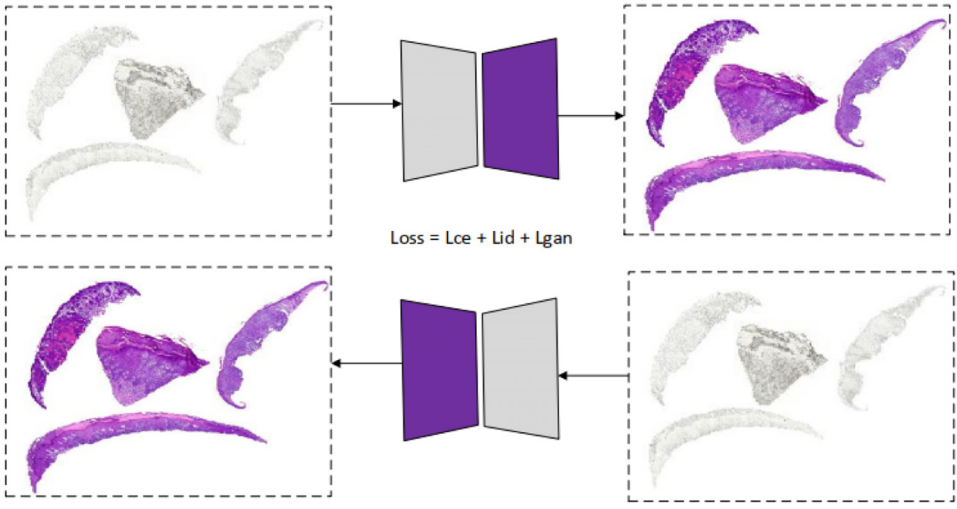

Fig. 2. Dual Contrastive GAN model trained on the acquired data repository to generate H&E-stained images from unstained images and vice versa.

Table 2
Evaluation of generated image compared with unstained and chemically stained images.

Image Type		FID Score
Chemically stained	Virtually stained	80.47
Chemically stained	Un-stained	320.4
Virtually stained	Un-stained	342.01

indicate a high degree of contextual correlation between virtually and chemically stained image samples, while showing substantial dissimilarity between stained and unstained images, thus validating the synthetic images. Furthermore, to validate the virtually stained images, a panel of experienced dermatopathologists assessed both conventional and digitally stained images for various quality factors such as color, resolution, sharpness, contrast, brightness, uniform illumination, artefacts, melanocytes, keratocytes, and inflammatory cells. This evaluation revealed an average concordance of 78.8 % and 90.2 % for assessments of paired and individual digitally stained images, respectively.

Limitations

Not applicable.

Ethics Statement

In conducting this study, we have taken rigorous steps to ensure ethical standards were upheld, particularly concerning the involvement of human participants. All participants provided informed consent, having been fully briefed on the nature of the study, the data being collected, its intended use, and any associated risks. To protect privacy, all personally identifiable information was removed from the dataset, and the data was anonymized to prevent direct or indirect identification of individuals.

Data Availability

E-Staining DermaRepo (Original data) (Mendeley Data).

CRediT Author Statement

Muhammad Zeeshan Asaf: Methodology, Software, Data curation, Writing – original draft; **Anum Abdul Salam:** Conceptualization, Software, Validation; **Samavia Khan:** Validation, Formal analysis; **Noah Musolff:** Writing – review & editing, Visualization; **Muhammad Usman Akram:** Conceptualization, Supervision, Project administration; **Babar Rao:** Data curation, Writing – review & editing, Resources, Supervision.

Acknowledgments

This research did not receive any particular grant from funding agencies in the public, commercial, or not-for-profit sectors. This declaration underscores the commitment to maintain integrity, transparency, and independence in the pursuit of scientific inquiry.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Chen, Y.T. Liu, F.S. Khan, M.C. Fox, J.S. Reichenberg, F.C. Lopes, K.R. Sebastian, M.K. Markey, J.W. Tunnell, (2024). Single color virtual H&E staining with In-and-Out Net. arXiv preprint arXiv:2405.13278.
[2] F. Chen, R. Zhang, B. Zheng, Y. Sun, J. He, W. Qin, (2024). Pathological semantics-preserving learning for H&E-to-IHC virtual staining. arXiv preprint arXiv:2407.03655.
[3] R.J. Hay, N.E. Johns, H.C. Williams, I.W. Bolliger, R.P. Dellavalle, D.J. Margolis, R. Marks, L. Naldi, M.A. Weinstock, S.K. Wulf, C. Michaud, The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions, J. Investig. Dermatol. 134 (6) (2014) 1527–1534.
[4] C. Karimkhani, R.P. Dellavalle, L.E. Coffeng, C. Flohr, R.J. Hay, S.M. Langan, E.O. Nsoesie, A.J. Ferrari, H.E. Erskine, J.I. Silverberg, T. Vos, Global skin disease morbidity and mortality: an update from the global burden of disease study 2013, JAMA Dermatol. 153 (5) (2017) 406–412.
[5] A.J. Ferrari, D.F. Santomauro, A. Aali, Y.H. Abate, C. Abbafati, H. Abbastabar, S. Abd ElHafeez, M. Abdelmasseh, S. Abd-Elsalam, A. Abdollahi, A. Abdullahi, Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021, Lancet 403 (10440) (2024) 2133–2161.
[6] C.R. Finley, D.S. Chan, S. Garrison, C. Korownyk, M.R. Kolber, S. Campbell, D.T. Eurich, A.J. Lindblad, B. Vandermeer, G.M. Allan, What are the most common conditions in primary care?: systematic review, Can. Fam. Physician 64 (11) (2018) 832–840.
[7] M.A. Richard, C. Paul, T. Nijsten, P. Gisondi, C. Salavastru, C. Taieb, M. Trakatelli, L. Puig, A. Stratigos, EADV Burden of Skin Diseases Project Team, Prevalence of most common skin diseases in Europe: a population-based study, J. Eur. Acad. Dermatol. Venereol. 36 (7) (2022) 1088–1096.
[8] M.Z. Asaf, B. Rao, M.U. Akram, S.G. Khawaja, S. Khan, T.M. Truong, P. Sekhon, I.J. Khan, M.S. Abbasi, Dual contrastive learning-based image-to-image translation of unstained skin tissue into virtually stained H&E images, Sci. Rep. 14 (1) (2024) 2335.