Data Article

Non-melanoma skin cancer segmentation for histopathology dataset

Simon M. Thomas\textsuperscript{a,b}, James G. Lefevre\textsuperscript{a}, Glenn Baxter\textsuperscript{b}, Nicholas A. Hamilton\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a} Institute for Molecular Bioscience, The University of Queensland, St Lucia, Australia  
\textsuperscript{b} MyLab Pathology, Salisbury, Australia

A R T I C L E   I N F O

Article history:  
Received 6 October 2021  
Revised 24 October 2021  
Accepted 15 November 2021  
Available online 19 November 2021

Keywords:  
Digital pathology  
Machine learning  
Medical imaging  
Image analysis

A B S T R A C T

Densely labelled segmentation data for digital pathology images is costly to produce but is invaluable to training effective machine learning models. We make available 290 hand-annotated histopathology tissue sections of the 3 most common skin cancers; basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and intraepidermal carcinoma (IEC). These non-melanoma skin cancers constitute over 90\% of all skin cancer diagnoses and hence this dataset gives an opportunity to the scientific community to benchmark analytic methodologies on a significant portion of the dermatopathology workflow. The data represents typical cases of the three cancer types (not requiring a differential diagnosis) across shave, punch and excision biopsy contexts. Each image is accompanied with a segmentation mask which characterizes the section into 12 tissue types, specifically: keratin, epidermis, papillary dermis, reticular dermis, hypodermis, inflammation, glands, hair follicles and background, as well as BCC, SCC and IEC. Included also are cancer margin measurements to work towards automated assessment of surgical margin clearance and tumour invasion. This leaves open many opportunities for researchers to utilize or extend the dataset, building upon recent work on image analysis problems in skin cancer (Thomas et al., 2021).

DOI of original article: 10.1016/j.compmedimag.2021.101998  
* Corresponding author.  
E-mail address: n.hamilton@uq.edu.au (N.A. Hamilton).  
Social media: 🌟 (N.A. Hamilton)

https://doi.org/10.1016/j.dib.2021.107587  
2352-3409/© 2021 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Specifications Table

| Subject | Computer Science, Artificial Intelligence, Computational Biology, Cancer Research, Pathology and Medical Technology |
|-----------------|---------------------------------------------------------------------------------------------------------------|
| Specific subject area | Semantic segmentation dataset for machine learning analysis of histopathology images of non-melanoma skin cancers. |
| Type of data | TIFF H&E images, PNG annotated images |
| How the data were acquired | The H&E slides were imaged using a DP27 Olympus microscope camera at 10x magnification with the condenser attached. The final images are a high-resolution mosaic stitched together using software available at https://github.com/smithomas-sci/HistolmageStitcher. The segmentation masks were created using the ImageJ software by painting over the relevant tissue type. The downsampled datasets were created in Python 3.6 using the OpenCV library v3.4.5. |
| Data format | Raw, Annotated |
| Parameters for data collection | Slides that had clear and unambiguous diagnostic features of basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and intra-epidermal carcinoma (IEC) were selected by a pathologist, and representative tissue sections were identified and imaged. Slides represented an age range of 34–96, with a median age of 70. Male and female proportions were ⅔ and ¼ respectively. |
| Description of data collection | H&E slides were provided by MyLab Pathology and imaged at 10x magnification to create high-resolution images of BCC, SCC, and IEC. With reference to these images, pixel-level segmentation masks were generated for 12 major tissue types, resulting in a full characterization of the tissue section/image. Combined the H&E and segmentation masks present input/output pairs for a variety of machine learning workflows. |
| Data source location | **Institution:** Institute for Molecular Bioscience - The University of Queensland  
**City/Town/Region:** Brisbane, Queensland  
**Country:** Australia |
| Data accessibility | Repository name: UQ eSpace  
Direct URL to data: 10.14264/8be4bd0 |
| Related research articles | [1] S.M. Thomas, J.G. Lefevre, G. Baxter, N.A. Hamilton, Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer, Medical Image Analysis. 68 (2021) 101915.  
[2] S.M. Thomas, J.G. Lefevre, G. Baxter, N.A. Hamilton, Characterization of Tissue Types in Basal Cell Carcinoma Images via Generative Modelling and Concept Vectors, Computerized Medical Imaging and Graphics. 94 (2021) 101998. |

Value of the Data

- High-quality hand-annotations are costly to produce and rare in current digital pathology repositories. This dataset represents over 250 h of manual annotation and curation, specifically for non-melanoma skin cancer, the most common form of skin cancer.
- This dataset enables researchers to compare and benchmark their results for a variety of tasks within the domain of skin cancer image analysis e.g. image segmentation, classification, margin-detection/measurement, specimen orientation and assessment of cancer invasiveness.
- The segmentations provide a full characterization of the tissue sections and so can be used for tasks outside of machine learning and may be of interest to other non-technical researchers. They also provide an excellent starting point for further annotation if researchers wish to include more tissue (sub) classes.
1. Data Description

The repository contains raw image data and associated segmentation masks. The 1x directory contains the images in original resolution, with 290 histology images in the Images directory in TIFF format, and their respective segmentation mask in the Masks directory, in PNG format. The images are named according to their diagnostic class and their number within it e.g. BCC_25.png. An example for the image and mask pairs is seen below in Fig. 1. The datasets are replicated for the 2x, 5x and 10x downsample factors, and can be found in their respective directories.

The color legend for the segmentations is described below in Table 1, and a picture can be found in the root of the repository, named 12_class_Palette.tif. The 12 classes correspond to 12 broad tissue types, namely: Glands (GLD), Inflammation (INF), Hair Follicles (FOL), Hypodermis (HYP), Reticular Dermis (RET), Papillary Dermis (PAP), Epidermis (EPI), Keratin (KER), Background (BKG), BCC, SCC, and IEC.

The MarginData directory provides (x,y) pixel coordinates for the margins of each specimen for the 10x dataset. The Images subdirectory of MarginData contains the original 10x images and the X subdirectory contains the corresponding mask. The y subdirectory contains CSV files, each with six points, designating the left- and right-most margins, as well as the deepest invasion point of the cancer. The points can be plotted over the 10x image and segmentation masks for further context.

The root directory also contains the names of files corresponding to the train, validation and test sets, used in Ref. [1]. They constitute a 70:15:15% split of the 290 images.

![Fig. 1. The original histology image of an excisional biopsy (left), with a resolution of 11,412 × 15,940 pixels (7.6 x 10.6 mm) and the corresponding 12 class segmentation mask (right). The mask provides a full characterization of the tissue section by allocating pixels to 1 of 12 broad tissue classes, including background. An example machine learning problem would be to learn the mapping from the input domain to the segmentation mask.](image)
Table 1
The 12 tissues classes, their codes and respective colors in the segmentation masks. The RGB values can be used to look-up the classes programatically.

| Tissue Type                        | Code | RGB Values | Color |
|------------------------------------|------|------------|-------|
| Glands (sebaceous and sweat)       | GLD  | 108, 0, 115|       |
| Inflammation                       | INF  | 145, 1, 122|       |
| Hair Follicle                      | FOL  | 216, 47, 148|     |
| Hypodermis                         | HYP  | 254, 246, 242|   |
| Reticular Dermis                   | RET  | 181, 9, 130|     |
| Papillary Dermis                   | PAP  | 236, 85, 157|    |
| Epidermis                          | EPI  | 73, 0, 106 |     |
| Keratin                            | KER  | 248, 123, 168|  |
| Background                         | BKG  | 0, 0, 0    |      |
| Basal Cell Carcinoma               | BCC  | 127, 255, 255|   |
| Squamous Cell Carcinoma            | SCC  | 127, 255, 142|   |
| Intra-epidermal Carcinoma          | IEC  | 255, 127, 127| |

2. Experimental Design, Materials and Methods

Non-melanoma skin cancers constitute over 90% of skin cancer diagnoses, with basal cell carcinoma and squamous cell carcinoma being the most common [3, 4]. Therefore, a dataset that enables the application of machine learning to the majority of skin cancer cases has significant value. MyLab Pathology provided access to their collection of histological skin cancer slides, which were processed using xylene and infused with paraffin wax. In consultation with a pathologist we selected 290 slides representing typical cases of BCC (140), SCC (60) and IEC (90). Cases where features were ambiguous and required a differential diagnosis were excluded. The data includes shave biopsies (100), punch biopsies (58) and excisional biopsies (132). For each slide, a pathologist hand-annotated the tissue section on the slide that was most representative of the diagnostic class. These tissue sections were then imaged over four months in late 2017 and early 2018. The resulting images correspond to patients within the age range of 34–96 with a median age of 70. The male and female proportions were 2/3 and 1/3, respectively, which closely reflects the prevalence of non-melanoma skin cancer in the Australian population [5].

The slides were imaged using a DP27 Olympus microscope camera at 10x magnification with the condenser attached. At this magnification, many images are required to capture the whole tissue section, and so overlapping tiles were combined into a high-resolution mosaic using bespoke software - https://github.com/smthomas-sci/HistoImageStitcher. The final resolution of the images ranged between 11 million and 500 million pixels, where 1 pixel corresponds to 0.67 μm in length.
The ground-truth segmentations were created by a trained histopathology scientist using the ImageJ software. Full-resolution images were painted over using colors indicating 1 of 12 classification categories (see Table 1). The epidermis presents a large amount of variation that can be non-cancerous, but departs considerably from healthy features. Therefore, the epidermis class contained only ideal health epidermis, and deviations from this were included in the IEC class, resulting in features ranging from mildly to severely dysplastic keratinocytes (solar keratosis) as well as carcinoma. This enabled multiple non-health features to be present in an image that did not fit into the same class. The hand-annotated segmentations took approximately 250 h to complete. The dataset was then downsample by factors of 2, 5 and 10 to enable a variety of experiments to be performed using features at different scales.

Ethics Statements

Consent was obtained from MyLab Pathology to use and publish the data collected in accordance with regulations. The Human Ethics Research Office at the University of Queensland assessed the data as non-identifiable data about human beings and of negligible risk under the National Statement on Ethical Conduct in Human Research (review number 2018001029).

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.

CRediT Author Statement

Simon M. Thomas: Conceptualization, Methodology, Software, Data curation, Writing – original draft; James G. Lefevre: Conceptualization, Supervision; Glenn Baxter: Conceptualization, Methodology, Data curation, Resources, Supervision; Nicholas A. Hamilton: Conceptualization, Supervision, Writing – review & editing, Resources, Project administration.

Acknowledgments

JL is supported by Australian Research Council Discovery Grant DP180101910. MyLab Pathology provided access to their archived histological collection. Financial support was provided by The Australian Government Research Training Program (RTP) and The Laurel Joy George Perpetual Scholarship.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107587.

References

[1] S.M. Thomas, J.G. Lefevre, G. Baxter, N.A. Hamilton, Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer, Med. Image Anal. 68 (2021) 101915.
[2] S.M. Thomas, J.G. Lefevre, G. Baxter, N.A. Hamilton, Characterization of tissue types in basal cell carcinoma images via generative modelling and concept vectors, Comput. Med. Imaging Graph. 94 (2021) 101998.
[3] A.C. Geller, G.D. Annas, Epidemiology of melanoma and nonmelanoma skin cancer, Semin. Oncol. Nurs. 19 (2003) 2–11.
[4] H.W. Rogers, M.A. Weinstock, S.R. Feldman, B.M. Coldiron, Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the U.S. population, 2012, JAMA Dermatol. 151 (2015) 1081–1086, doi:10.1001/jamadermatol.2015.1187.

[5] M.P. Staples, M. Elwood, R.C. Burton, J.L. Williams, R. Marks, G.G. Giles, Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985, Med. J. Aust. 184 (2006) 6–10.