International Journal of Research and Review
DOI: https://doi.org/10.52403/ijrr.20211034
Vol.8; Issue: 10; October 2021
Website: www.ijrrjournal.com
Original Research Article *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

A Collection of Multimodal Melanoma Data

Aigli Korfiati[1], Giorgos Livanos[1], Christos Konstandinou[1], Sophia Georgiou[2], George Sakellaropoulos[1]

[1]Department of Medical Physics, School of Medicine, University of Patras, Greece
[2]Department of Dermatology, School of Medicine, University of Patras, Greece

Corresponding Author: George Sakellaropoulos

ABSTRACT

Computer-aided diagnosis, prognosis and therapy systems have been of great interest for a number of years. The availability of big volumes of data and of powerful computational resources have allowed artificial intelligence approaches to emerge in melanoma related studies. However, for such approaches to have good predictive performances data availability is of crucial importance. Melanoma related imaging, biological and clinical data can be found partially and scattered in various repositories. Thus, in this work, we assemble in a web accessible database, named ebioMelDB, the widest collection of clinical and dermoscopy images accompanied with patient clinical data and the widest collection of RNA-Seq gene expression data accompanied with patient clinical data. The database organization allows users to select the data that are appropriate for their application of interest (diagnosis, prognosis and therapy).

Key Words: melanoma database, integrated data, dermoscopy, imaging, RNA-Seq, clinical data

INTRODUCTION

Cutaneous melanoma is a rare skin cancer, but is associated with high mortality rates [1]. Thus, the research community has shown interest in building automated frameworks to support its diagnosis, prognosis and therapy, starting from 1985 when the ABCD rule [2] was devised by Kopf et al. to guide clinicians in the detection of melanoma. More sophisticated computer aided diagnosis (CAD) systems using dermoscopy or clinical images started subsequently to emerge [3]. However, apart from examining the lesions, biological, demographic and lifestyle variability of the patients has to be taken into account [4] when it comes to a multi-factorial disease, like melanoma.

For computer systems to succeed in aiding disease diagnosis, prognosis and treatment, data availability is a necessity. Imaging, clinical and biological data streams have to be collected, harmonized, integrated and analyzed before being employed in CAD systems. Additionally, the artificial intelligence algorithms used in CAD systems depend on data availability [5], since they learn from the data in order to be able to perform predictions.

In this context, ebioMelDB [6] was generated as a melanoma-specific database of clinical, biological and imaging data. In the present work, we describe an update of ebioMelDB with 2500 additional image entries and 468 additional biological data entries. In its current version, all publicly available clinical images, dermoscopy images and RNA-Seq gene expression data, integrated with clinical data have been gathered in one repository, which can be accessed through http://www.med.upatras.gr/ebioMelDB. Manual data curation and organization in categories allows for the selection of the most appropriate data subset for the user's application of interest (diagnosis, prognosis and therapy).

IMAGE DATA COLLECTION

We consolidated publicly available for academic and research purposes dermoscopy and clinical images. The majority of the dermoscopic data is part of the International Skin Imaging Collaboration (ISIC) Archive, a collection of dermoscopy images which was released during the challenges hosted from ISIC organization from 2016 to 2020. It consists of dermoscopy images along with patients' clinical data. In more detail, ISIC2016 [7] Task 3, ISIC2017 [8] Task 3 and ISIC2018 [9] Task 3 training data were excluded, since they are part of ISIC2019 [8,10,11] training data, which, along with the ISIC 2020 Challenge [12] training data are included in ebioMelDB. Another source of dermoscopy images is the PH² database [13]. The 7-pt dataset [14], which is a collection of dermoscopy and clinical images complemented with patient clinical data is also incorporated in ebioMelDB. Another source of clinical images is MED-NODE [15]. This latest version of ebioMelDB includes 220 additional dermoscopy images, retrieved from the ISIC2018 Validation set of images and 2298 clinical images from the recently published PAD-UFES-20 [16] dataset which contains clinical images accompanied with clinical information.

With this update, the total count of images incorporated in ebioMelDB raises from 63480 to 65980. The number of images retrieved from each distinct database and the number of images hosted in ebioMelDB are presented in Table 1. Among them 62528 are dermoscopy and 3470 are clinical images. After careful examination of the images in each source dataset, duplicate images were observed and removed from ebioMelDB. The numbers after duplicates removal are also presented in Table 1. Dermoscopy images originate from ISIC2018, ISIC2019, ISIC2020, PH² and 7-pt. Clinical images originate from 7-pt, MED-NODE and PAD-UFES-20. Clinical data accompany and are available for some of the images. In more detail, the patient's sex is available for 61114 images, the patient's age for 59849 images and the lesion's anatomical site for 59186 images.

Table 1: Number of images hosted in different databases.

Databases	7-pt	isic2018	isic2019	isic2020	mednode	ph2	padufes	ebioMelDB
Number of images	2022	202	27962	33126	170	200	2298	65980
Number of images – without duplicates	2013	193	25331	32701	170	200	2298	62906

Table 2: The distribution of dermoscopy images in 5 benign and malignant categories. The ISIC2020 contains a portion of data that is classified as unknown; thus, it is not included in the 5-class classification.

	isic18	isic19	isic20	ph2	7pt	ebioMelDB
MEL	21	4522	581	40	252	5986
NV	132	12875	5191	80	572	19737
NMC	15	3951	0	0	42	4709
NNV	26	3116	222	0	145	3833
SUS	8	867	1	80	0	1114

Table 3: The distribution of clinical images in 5 benign and malignant categories.

	padufes	mednode	7pt	ebioMelDB
MEL	52	70	249	371
NV	244	100	575	919
NMC	1037	0	42	1079
NNV	235	0	136	371
SUS	730	0	0	730

The different source databases provide the diagnosis of the images in different levels of generalization and with different naming conventions. To address the diversity of the datasets in terms of diagnosis, we integrated them into five categories: NV (Nevus), MEL (Melanoma), NNV (Non-Nevus benign), NMC (Non-Melanocytic Carcinoma) and SUS (Benign-Suspicious). The benign categories are NV, NNV and SUS. The nevus category NV includes images that have been diagnosed as simple, common, blue, clark, combined, con-genital, dermal, recurrent, reed or spitz nevus. The benign non-nevus category NNV includes images that have been diagnosed as dermatofibroma, lentigo, melanosis, miscellaneous, seborrheic keratosis, vascular lesion, benign keratosis, cafe-au-lait macule, lentigo NOS and lichenoid keratosis. The benign, but suspicious for malignancy category SUS includes images

that have been diagnosed as actinic keratosis, atypical melanocytic proliferation and atypical nevus. The malignant melanoma category MEL includes melanoma and melanoma metastases images. Finally, the category of non-melanocytic carcinomas includes images with basal cell carcinoma and squamous cell carcinoma. In Table 2 and Table 3 we present the images distribution for each image type (dermoscopy and clinical) in the aforementioned five broader categories.

The mapping between these classes and the initial provided labels in each distinct source database is displayed in Table 4.

Table 4: Grouping of the images diagnosis naming conventions as presented in each source database into 5 broader categories: Nevus (NV), Benign Non-Nevus (NNV), Benign but Suspicious for malignancy (SUS), Melanoma (MEL) and Non-Melanocytic Carcinomas (NMC).

	ebioMelDB	isic18	isic19	isic20	7pt	ph2	mednode	padufes
Benign	**NV**	NV	NV	nevus	nevus clark, combined nevus, blue nevus, congenital nevus, dermal nevus, recurrent nevus, reed or spitz nevus	Common Nevus	naevus	NEV
	NNV	BKL DF VASC	BKL DF VASC	seborrheic keratosis, lichenoid keratosis, lentigo NOS, solar lentigo, cafe-au-lait macule	dermatofibroma, melanosis, lentigo, seborrheic keratosis, vascular lesion, miscellaneous	—	—	SEK
	SUS	AKIEC	AK	atypical melanocytic proliferation	—	Atypical Nevus		ACK
Malignant	**MEL**	MEL	MEL	melanoma	melanoma, melanoma metastasis	Melanoma	melanoma	MEL
	NMC	BCC	BCC, SCC	—	basal cell carcinoma	—	—	BCC, SCC

BIOLOGICAL DATA COLLECTION

Biological data in ebioMelDB are RNA-Seq gene expression data related to melanoma focusing only on the human organism. With this update of ebioMelDB, the total number of biological samples raises from 4490 to 4958.

For the biological data collection, we searched the public databases NCBI Gene Expression Omnibus (GEO) database [17] and The Cancer Genome Atlas (TCGA) [18]. The GEO (http://www.ncbi.nlm.nih.gov/geo/) repository hosts high-throughput microarray and next-generation sequencing functional genomic data sets submitted by the research community. GEO data include raw data, processed data and metadata and are organized in series (GSE) of datasets, which are groups of samples. The R package GEOmetadb [19] was used to download the data and their related metadata. In specific, the keyword "melanoma" was searched against the following GEO fields: GSE titles, summaries and overall designs. From the resulting data series, we kept only those with Expression profiling by high throughput sequencing as experiment type and Homo sapiens as organism. This resulted in 291 series, which were subsequently manually curated to keep only series that actually included melanoma datasets, ending up with 178 series that consist of 4490 samples. TCGA is a repository containing genomic, transcriptomic, epigenetic, proteomics and clinical information of 33 types of cancer, among which melanoma with its TCGA-SKCM project [20]. The TCGA-SKCM project includes 468 samples of RNA-Seq gene expression data.

In order to better assist researchers find the most appropriate data subset for their application of interest the data was organized in a number of potentially overlapping categories. For diagnosis related studies, healthy control or data from other non-melanoma diseases are required. So, the first category is the presence or not of healthy control, non-melanoma samples and the second is whether samples of another disease are present in the data

series. The third category clinical information indicates whether accompanying clinical information, such as age, sex, disease state, vital status, etc. is available for the data series samples. Such data can be used for both diagnostic and prognostic purposes. The fourth category is treatment and indicates that some samples of the data series were treated with a specific drug or other kind of treatment. Researchers interested in treatment studies can find appropriate data in the treatment category or in the variation category. This fifth category variation indicates various types of perturbations in the samples, such as the overexpression or knockdown of a gene (which could be used as drug targets or help us better understand the disease mechanism), or resistance to a therapy. Finally, a group of four categories is related to the biological origin of the samples and includes patients' specimens, cell lines, xenograft models and other cells. For each category assigned to a data series, there is also a text field with a brief description of why the category is assigned. The numbers of data series and sample in each of the aforementioned overlapping categories is presented in Table 5.

Table 5: Numbers of biological data series and samples in ebioMelDB categories.

Category	Number of data series	Number of samples
control	28	726
other disease	14	546
clinical information	17	991
treatment	75	2370
variation	91	2112
patients	32	1434
cell lines	124	3163
xenograft models	13	265
other cells	29	727
Total	179	4958

DATABASE ORGANIZATION

EbioMelDB is an online database that can be accessed through http://www.med.upatras.gr/ebioMelDB hosting image and biological data for melanoma, accompanied with patient clinical information. One page hosts the images and another one the biological data. Each biological series or image can be viewed in more detail in a dedicated view page. For each image, the user can view the actual image, the initial diagnosis, the image category in the 5 proposed broader categories, the image type (dermoscopy or clinical) and accompanying clinical data, like patient's sex, age and lesion's anatomical site among others. For each biological data series, the user can view its title, its ftp download link, its publication, submission and last update dates, its summary and overall design, contributors and contact person, the name of the sequencer the samples were generated with and the total number of samples in the series. The user can also view whether the series belongs or not to the aforementioned 9 application-specific or biological origin-specific categories.

The images and the biological data are organized in datatables which are searchable and sortable for better browsing. Additionally, in order to allow users access in a more targeted way the data of their interest, filtering based on user defined criteria is available. Filtering can be performed both on the data categories and characteristics. For example, a user interested in series of biological data series with a) patients' specimen, that also have b) clinical info, c) control samples and d) the samples count of each series is 20 or more, can apply the relevant filters and retrieve only 4 out of the 179 data series hosted in ebioMelDB.

For the implementation of the online database, the Django Web Framework together with python3.8, html5, JavaScript and SQLite were used. EbioMelDB is running on a Linux server.

CONCLUSION

Computer-aided diagnosis, prognosis and therapy systems for melanoma have been of great interest for a number of years, but only recently with the availability of big volumes of data and of powerful computational resources, artificial intelligence approaches were enabled to emerge in melanoma related studies.

However, for artificial intelligence approaches to have good predictive performances, data availability is of crucial importance. Melanoma related data that could be used for diagnostic, prognostic and therapy purposes include imaging, biological and clinical data. Currently, public such data can be found partially and scattered in various repositories.

Thus, in this work, we assembled in a web accessible database, named ebioMelDB, the widest collection of clinical and dermoscopy images accompanied with patient clinical data and the widest collection of RNA-Seq gene expression data accompanied with patient clinical data. EbioMelDB hosts 65980 clinical and dermoscopy images organized in 3 benign and 2 malignant categories. It also hosts 4958 human RNA-Seq gene expression melanoma samples. The database organization allows users to select the subset of data that is most appropriate for their application of interest (diagnosis, prognosis and therapy).

ACKNOWLEDGEMENT

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning 2014-2020» in the context of the project "Biomarkers extraction from imaging and molecular biology data using computational models to assist malignant melanoma diagnosis, prognosis and treatment" (MIS 5047174).

Conflict of Interest: None

Source of Funding: None

REFERENCES

1. Rebecca VW, Somasundaram R, Herlyn M. Pre-clinical modeling of cutaneous melanoma. Nature communications. 2020 Jun 5;11(1):1-9.
2. Ali AR, Li J, Yang G. Automating the ABCD rule for melanoma detection: a survey. IEEE Access. 2020 Apr 28;8:83333-46.
3. Barata C, Celebi ME, Marques JS. A survey of feature extraction in dermoscopy image analysis of skin cancer. IEEE journal of biomedical and health informatics. 2018 Jun 11;23(3):1096-109.
4. Dimitriou F, Krattinger R, Ramelyte E, Barysch MJ, Micaletto S, Dummer R, Goldinger SM. The world of melanoma: epidemiologic, genetic, and anatomic differences of melanoma across the globe. Current oncology reports. 2018 Nov;20(11):1-9.
5. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. Computers in Biology and Medicine. 2020 Oct 27:104065.
6. Korfiati A, Livanos G, Konstantinou C, Georgiou S, Sakellaropoulos G. ebioMelDB: Multi-modal Database for Melanoma and Its Application on Estimating Patient Prognosis. InIFIP International Conference on Artificial Intelligence Applications and Innovations 2021 Jun 25 (pp. 33-44). Springer, Cham.
7. Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1605.01397. 2016 May 4.
8. Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H, Halpern A. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) 2018 Apr 4 (pp. 168-172). IEEE.
9. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, Helba B, Kalloo A, Liopyris K, Marchetti M, Kittler H. Skin lesion analysis toward melanoma

detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368. 2019 Feb 9.
10. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data. 2018 Aug 14;5(1):1-9.
11. Combalia M, Codella NC, Rotemberg V, Helba B, Vilaplana V, Reiter O, Carrera C, Barreiro A, Halpern AC, Puig S, Malvehy J. BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288. 2019 Aug 6.
12. Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, Combalia M, Dusza S, Guitera P, Gutman D, Halpern A. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data. 2021 Jan 28;8(1):1-8.
13. Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH 2-A dermoscopic image database for research and benchmarking. In2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2013 Jul 3 (pp. 5437-5440). IEEE.
14. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics. 2018 Apr 9;23(2):538-46.
15. Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert systems with applications. 2015 Nov 1;42(19):6578-85.
16. Pacheco AG, Lima GR, Salomão AS, Krohling B, Biral IP, de Angelo GG, Alves Jr FC, Esgario JG, Simora AC, Castro PB, Rodrigues FB. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in brief. 2020 Oct 1;32:106221.
17. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A. NCBI GEO: archive for functional genomics data sets- update. Nucleic acids research. 2012 Nov 26; 41(D1):D991-5.
18. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology. 2015;19(1A):A68.
19. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics. 2008 Dec 1;24(23):2798-800.
20. Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB, Arachchi H, Arora A, Auman JT, Ayala B, Baboud J. Genomic classification of cutaneous melanoma. Cell. 2015 Jun 18;161(7):1681-96.

How to cite this article: Korfiati A, Livanos G, Konstandinou C et.al. A collection of multimodal melanoma data. *International Journal of Research and Review*. 2021; 8(10): 257-262. DOI: *https://doi.org/10.52403/ijrr.20211034*

******