ebioMelDB: Multi-modal Database for Melanoma and Its Application on Estimating Patient Prognosis

Aigli Korfiati, Giorgos Livanos, Christos Konstantinou, Sophia Georgiou, George Sakellaropoulos

ebioMelDB: Multi-modal database for melanoma and its application on estimating patient prognosis

Aigli Korfiati[1], Giorgos Livanos[1], Christos Konstantinou[1], Sophia Georgiou[2] and 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
aigi.korfiati@gmail.com, livanosg@upnet.gr, chrikon@upatras.gr, sgeo@upatras.gr, gsak@upatras.gr

Abstract. Data availability is important when researchers want to apply artificial intelligence algorithms to extract biomarkers and generate predictive models for disease diagnosis, response to treatment and prognosis. For cutaneous melanoma clinical, biological and imaging data are scattered through the web. ebioMelDB is the first database to integrate the widest collections of RNA-Seq gene expression and clinical data with clinical and dermoscopy images, all manually curated and organized in categories. ebioMelDB aspires also to host our under development predictive models in cutaneous melanoma diagnosis, response to treatment and prognosis based on combinations of the different data types hosted. As a first step towards this direction, we apply an ensemble dimensionality reduction technique employing a multi-objective optimization heuristic algorithm that finds the best feature subset, the best classifier among linear SVM, Radial Basis Function Kernel SVM and random forest and their optimal parameters to predict the vital status of patients in different time windows based on a large cohort of patients' gene expression data. The results are very encouraging in performance metrics compared with state-of-the-art algorithms. The database is available at http://www.med.upatras.gr/ebioMelDB.

Keywords: Cutaneous Melanoma, Database, Prognosis, Machine learning.

1 Introduction

Cutaneous melanoma (CM), commonly developed from malignant transformation of melanocytes cells that produce melanin in the skin, constitutes ~5% of all skin cancers [1]. However, >75% of skin cancer deaths originate from melanoma, which has a 5-year survival rate of 23% in patients with late stage of the disease [1]. Early detection is the most important determinant of the associated mortality reduction and towards this direction, in 1985, the ABCD rule [2] was devised by Kopf et al. as a simple framework that physicians, novice dermatologists and non-physicians can use to detect melanocytic lesions with atypical features. Based on the above rule, atypical melanocytic nevi are characterized by A (Asymmetry), B (Border irregularities), C (Colors variety) and D (Diameter > 6mm) [2].

The introduction and application of dermoscopy in clinical practice has provided a new dimension in the evaluation of pigmented skin lesions. With this non-invasive technique and the pattern analysis method, melanocytic lesions are distinguished from non-melanocytic lesions. However, dermoscopy as a method has its limitations, as it depends on the examiner's experience. Traditional radiomics practice uses machine learning methods towards the development of computer-aided diagnosis (CAD) tools that can be used by dermatologists to overcome the aforementioned issues [3]. These systems follow a pipeline: i) image preprocessing, ii) lesion segmentation, iii) feature extraction, iv) feature selection (optional), and v) classification. Recently, a vast number of deep learning methods have been employed in CM research, but these as well have challenges and limitations [4].

CM is considered a multifactorial disease, the result of genetic predisposition and environmental factors [5] and survival outcomes and response to treatment can vary widely among patients due to the biological heterogeneity of melanoma [6]. Thus, in the effort to better understand the disease mechanisms and apply individualized treatment protocols to CM patients, omics data have been explored for the identification of diagnostic and prognostic biomarkers. Four subtypes of cancer, which include mutant BRAF, mutant RAS, mutant NF1, triple WT (wild-type) based on mutant genes have been widely reported [7] and a recent review on the genomic features characterizing the development of CM can be found in [8]. Apart from genomics, studies have demonstrated that CM arises from the anomalies in transcriptomic and epigenetic factors such as expression of mRNAs, miRNAs, the aberration in methylation patterns of CpG islands of genes and histone modifications [9]. In an attempt to find melanoma subtypes based on transcriptomics data, four major signatures have been found: immune, keratin, melanocyte inducing transcription factor (MITF)-low and MITF-high [10]. In the last decade, an extraordinary leap forward in the treatment of melanoma occurred, taking advantage of the advent of targeted therapies and immunotherapies [11]. At present, the methods commonly used in the treatment of melanoma include surgical resection, chemotherapy and immunotherapy. Interestingly, 13 FDA-approved treatments are presented in a review by Donelly et al [12] for different molecular subtypes of the disease, most originating from CM omics biomarkers. But again, because of the molecular heterogeneity, not all patients respond well to treatments and some present drug resistance. Therefore, it is imperative to develop prognostic biomarkers for risk stratification and treatment optimization [13].

However, analyzing only a single type of omics measurement poses limitations, because it cannot comprehensively and accurately describe the biological processes underlying the disease and may lead to partial and uninformative biomarkers. Thus, multidimensional studies which profile multiple types of omics changes on the same subjects have emerged [14]. A representative example is TCGA (The Cancer Genome Atlas) which is organized by NIH. TCGA is one of the most prominent and inclusive repositories containing genomic, transcriptomic, epigenetic, proteomics and clinical information of 33 types of cancer [15], among which CM with its TCGA-SKCM project [7].

Several studies have been conducted to identify prognostic biomarkers based on various TCGA-SKCM omics data with most of them including expression data. Jiang et al. [14] focused on a multi-omics analysis by integration of mutation, copy number variation, methylation, and messenger RNA expression data to achieve this objective, while the authors in [16] identified molecular subtypes associated with differences in CM prognosis by integrating epigenomic and genomic data. The authors in [17] support that integrating gene expression regulators when analyzing gene expression data can more accurately identify biomarkers. A number of studies have generated immune related prognostic gene signatures (a 239-gene signature in [18], a 33-gene signature in [19], a 25-gene signature in [20], a 7-gene signature in [21] and a 6-gene signature in [22]). In [23] the authors generated a 121 metastasis-associated prognostic signature and the authors in [9] are trying to distinguish metastatic melanoma from primary tumors based on the mRNA, miRNA and methylation data of TCGA providing their prediction models through the webserver, CancerSPP. The STATegra framework is presented in [24] as a multi-omics integrative pipeline used on mRNA and miRNA expression and methylation data. Several studies present web-servers which provide survival analysis based on gene expression [13]. However, these tools analyze statistically the association between single genes and survival prognosis in TCGA cancers.

Integrating omics and clinical data with imaging data is promising [25]. In melanoma, there is limited relevant research. In [26], the authors studied melanoma prognosis in terms of recurrence-free survival, based on clinical data, gene expression, and whole slide image features. Their best performing model included 20 automatically generated whole slide image features, 3 clinicopathologic variables, and mutation status of 2 genes. Maglogiannis et al [27] propose a platform able to integrate omics, histological images and clinical data for skin cancer patients and construct a synthetic dataset with mutated genes and images in order to discriminate melanoma from dysplastic nevi.

In the present paper we introduce ebioMelDB, a multi-modal database for cutaneous melanoma aspiring to enable researchers perform studies for extraction of biomarkers combining different types of data, including clinical, biological and imaging data. In its current version, ebioMelDB incorporates publicly available RNA-Seq gene expression data from GEO and TCGA, manually curated and organized in categories. It also includes the widest collection of clinical and dermoscopy images organized in 3 benign and 2 malignant categories, including Nevus, Benign Non-Nevus, Benign but Suspicious for malignancy, Melanoma and Non-Melanocytic Carcinomas, respectively. In its future versions, ebioMelDB will host our predictive models in CM diagnosis, response to treatment and prognosis based on combinations of the different data types hosted. As a first step towards this vision, we apply a machine learning classifier to predict the vital status of CM patients in different time windows based on gene expression data from TCGA.

2 Database

2.1 Image data collection

The image data is a collection of diverse public datasets and include:

- The dataset provided from Kawara et al. [28] (referred herein as 7-PT), which has been used for 7-point melanoma checklist criteria classification and skin lesion diagnosis, including 2022 dermoscopic and clinical images of the lesions.
- 27962 images from ISIC 2019 [29-31] Challenge dataset for dermoscopic image classification among nine different diagnostic categories.
- 33126 images from ISIC 2020 [32] Challenge for dermoscopic image classification tasks of benign and malignant skin lesions.
- 170 non-dermoscopic image dataset from Giotis et al. [33] on the computer-assisted diagnostic system MED-NODE.
- The PH^2 [34] dermoscopic image dataset which contains 200 images for common/atypical nevi and melanoma.

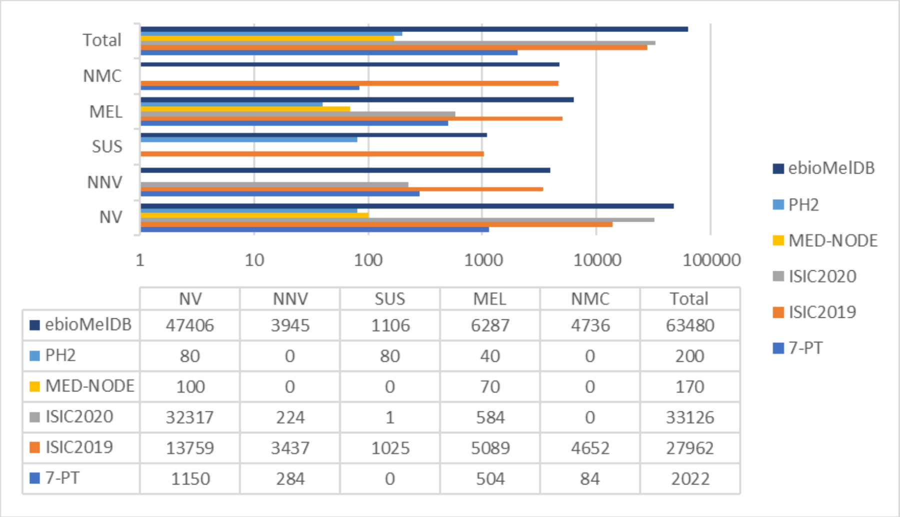
The different datasets include different naming conventions and different skin disease categories. Due to this high diversity, it was necessary to merge the diagnostic classes under broader categories. The broader categories we selected are 3 benign categories including Nevus (NV), Benign Non-Nevus (NNV), Benign but Suspicious for malignancy (SUS) and 2 malignant categories including Melanoma (MEL) and Non-Melanocytic Carcinomas (NMC). The grouping of the naming conventions of the different datasets is presented in Table 1.

Table 1. Naming conventions as presented in the original datasets and the respective grouping in ebioMelDB categories: Nevus (NV), Benign Non-Nevus (NNV), Benign but Suspicious for malignancy (SUS), Melanoma (MEL) and Non-Melanocytic Carcinomas (NMC)

Categories	NV	NNV	SUS	MEL	NMC
7-PT	blue, clark, combined, congenital, dermal, recurrent, reed or spitz nevus	dermatofibroma, lentigo, melanosis, miscellaneous, seborrheic keratosis, vascular lesion,	-	melanoma, melanoma metastasis	basal cell carcinoma
ISIC2019	NV	BKL, DF, VASC	AK	MEL	BCC, SCC
ISIC2020	nevus, unknown	cafe-au-lait macule, lentigo NOS, lichenoid keratosis	atypical melanocytic proliferation	melanoma	-
MED-NODE	naevus	-	-	melanoma	-
PH^2	Common Nevus	-	Atypical Nevus	Melanoma	-

Figure 1 presents the counts of images originating from each dataset when assigned to ebioMelDB categories.

Information like image, image type and diagnosis exist for images of all datasets. The next most frequent fields are the anatomical site (88.6%), sex (96.7%) and age (93.5%). The counts of dermoscopic and clinical images are 62308 and 1172, respectively.



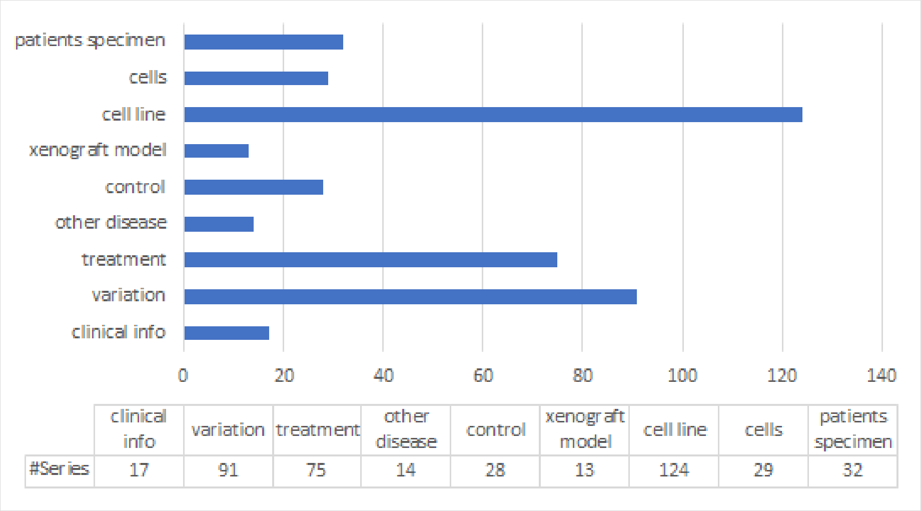
	NV	NNV	SUS	MEL	NMC	Total
■ ebioMelDB	47406	3945	1106	6287	4736	63480
■ PH2	80	0	80	40	0	200
■ MED-NODE	100	0	0	70	0	170
■ ISIC2020	32317	224	1	584	0	33126
■ ISIC2019	13759	3437	1025	5089	4652	27962
■ 7-PT	1150	284	0	504	84	2022

Fig. 1. Image distribution along datasets and ebioMelDB categories.

2.2 Biological data collection

Biological data were collected from the NCBI Gene Expression Omnibus (GEO) [35]. GEO (http://www.ncbi.nlm.nih.gov/geo/) is a public repository for high-throughput microarray and next-generation sequencing functional genomic data sets submitted by the research community including raw data, processed data and metadata and organized in series (GSE) of datasets. To download the data and their related metadata, the R package GEOmetadb [36] was used. The keyword "melanoma" was searched against all the GSE titles, summaries and overall designs, selecting only 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.

In order to better organize the data, we characterized them as belonging or not to a number of categories. The first group of categories is related to the origin of the biological samples and includes patients' specimens, cell lines, xenograft models and other cells. Another category is the presence or not of healthy control, non-melanoma samples to facilitate users aspiring to perform diagnostic studies. For the same reason, the category other disease indicates whether samples of another disease exist in a series. The category treatment shows that some samples of the series were treated with a specific drug or other kind of treatment to facilitate users interested in treatment studies. The category variation includes 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, the category clinical information indicates whether accompanying clinical information, such as age, sex, disease state, vital status, etc. is available for the samples. For each category assigned to a series, there is also a respective field with a brief description of why the category is assigned. Summary statistics of the database series of datasets are presented in Fig. 2.



	clinical info	variation	treatment	other disease	control	xenograft model	cell line	cells	patients specimen
#Series	17	91	75	14	28	13	124	29	32

Fig. 2. Number of series of RNA-Seq gene expression datasets belonging to the defined categories.

2.3 Database infrastructure

All the collected images and biological data are organized in a web accessible database at http://www.med.upatras.gr/ebioMelDB. One page presents the images and another one the biological data. Both are organized in datatables which are searchable and sortable. Moreover, filtering based on user defined criteria on the data categories and characteristics is available enabling the user to access in a more targeted way data of interest. For example, for the biological data, if a user is interested in 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, he applies the relevant filters and only 3 out of the 178 series are presented. Each biological series or image can be viewed in more detail in a dedicated view page. The database is developed based on the Django Web Framework, with the usage of python3.8, html5, JavaScript and SQLite and is running on a Linux server.

3 Estimating Melanoma Prognosis

3.1 Data Collection and Preprocessing

Gene expression quantification data were downloaded from the GDC Data Portal [37] of NIH. GDC hosts among others TCGA (https://www.cancer.gov/tcga) data, that has molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. For each cancer type, a combination of molecular biology (mRNA, protein and miRNA expression, copy number, DNA, DNA methylation), clinical and whole slide images data for the same patients are provided. The TCGA-SKCM [7] project has 470 cases of patients and data ranging the following categories: simple nucleotide variation, copy number variation, transcriptome profiling, biospecimen, sequencing reads, DNA methylation and clinical. For the estimation of melanoma prognosis, we downloaded the RNA-Seq gene expression quantification files (HTSeq-Counts) with a total of 472 files for 468 cases. The respective clinical data were also downloaded. 247 of them had vital status alive, 224 dead and 1 case with not reported vital status was excluded from the analysis. Custom python scripts were created to process the single count files and merge them in one file with 471 samples (expression files) and 60488 features (genes). The vital status was also mined from the clinical data and matched to the samples as labels with python scripts.

Genes with count less than 5 reads per sample were excluded from further analysis narrowing the number of genes down to 11339. Count data normalization and statistical analysis was performed with InSyBio Biomarkers tool (https://www.insybio.com/biomarkers.html). For the statistical analysis, the wilcoxon ranked sum test was employed and correction of p-values for multiple testing was performed using the Benjamini-Hochberg FDR adjustment method. Setting the adjusted p value threshold to 0.01, we identified 611 statistically significant differentially expressed genes.

3.2 Machine Learning Algorithm Description

In order to predict the patients' vital status from the gene expression data, we applied the machine learning method incorporated in the InSyBio Biomarkers tool. This method is an extension of the one presented in [38,39] and is an ensemble dimensionality reduction technique employing a multi-objective optimization heuristic algorithm that finds the best feature subset, the best classifier among linear SVM, Radial Basis Function Kernel SVM and random forest and their optimal parameters. In this extended version, multiple models performing equally well on the user-defined goals are the final outcome. And the final prediction is the one made by the majority of the classifiers. The weights used for the goals were Selected Features Number Minimization 1, Accuracy 10, F1 score 10, F2 score 1, Precision 1, Recall 1, ROC_AUC 1, Number of SVs or Trees Minimization 1.

3.3 Results

The experiments for the prediction of the patients' vital status were performed with 5-fold cross validation and for 100 generations. The cross-validation accuracy achieved was 68.84% with specificity 70.61% and sensitivity 67.22%. The selected features were 414 and the predictions were based on 25 Random Forest models, each with different number of trees ranging from 10 to 471.

The vital status is defined by TCGA as the survival state of the person and can have values dead or alive. The time period it refers to is up to almost 30 years after the diagnosis (10863 days). Thus, we wanted to examine the vital status of the patients in a 5-year, 3-year and 1-year period. From the clinical TCGA data and the Days to death field we computed the respective labels treating again each problem as a two-class classification problem. The respective 5-fold cross validation metrics (accuracy, specificity and sensitivity), the number of selected features, the number of classification models and their characteristics are presented in Table 2. The performance metrics get better as the examined time slot gets shorter.

Table 2. 5-fold cross validation metrics (accuracy -ACC, specificity-SP and sensitivity-SEN), number of selected features, number of classification models and their characteristics in the prediction of total, 5-,3- and 1-year vital status.

	vital status (in a ~30-year follow-up)	5-year vital status	3-year vital status	1-year vital status
Cross validation ACC	68.84 %	92.46 %	97.80 %	100.00 %
Cross validation SP	70.61 %	99.69 %	99.73 %	100.00 %
Cross validation SEN	67.22 %	85.23 %	95.88 %	100.00 %
# Selected features	414	13	163	7
# Classification models	25	11	16	4
# RF models	25 with 10-471 trees	1 with 38 trees	7 with 333-352 trees	-
# SVM models	-	10 rbf SVM with 435-471 SVs	9 rbf SVM with 471 SVs	4 rbf SVM with 282-459 SVs

Table 3. Comparison of the proposed method in terms of accuracy -ACC, specificity-SP and sensitivity-SEN in the prediction of total, 5-,3- and 1-year vital status.

	vital status (in a ~30-year follow-up)			5-year vital status		
metrics	ACC	SP	SEN	ACC	SP	SEN
WEKA SVM	54.14%	59.50%	48.20%	60.08%	72.10%	33.80%
WEKA RF	59.45%	63.20%	55.40%	67.94%	98.50%	1.40%
proposed method	**68.84%**	**70.61%**	**67.22%**	**92.46%**	**99.69%**	**85.23%**
	3-year vital status			1-year vital status		
metrics	ACC	SP	SEN	ACC	SP	SEN
WEKA SVM	65.45%	79.20%	24.50%	91.08%	96.20%	4.80%
WEKA RF	76.43%	99.70%	0.00%	94.48%	100.00%	0.00%
proposed method	**97.80%**	**99.73%**	**95.88%**	**100.00%**	**100.00%**	**100.00%**

Next, we wanted to compare with other methods, so we employed the SVM and Random Forest implementations of WEKA with default parameters and with 5-fold cross validation and the results are shown in Table 3. The proposed method clearly outperforms the other two methods in all cases, and the fact that the proposed method handles better imbalanced datasets is even more clear in the 5-, 3- and 1-year prediction problems where the samples of the minority class are 148, 110 and 27, respectively.

4 Discussion

CM is a skin cancer with high mortality and although diagnosis and treatment methods have made progress, its survival rate is still poor. Many studies, taking advantage of multi-omics data available in repositories like TCGA, have identified several prognostic biomarkers for CM. Identifying prognostic omics markers leads to a better understanding of the biological mechanisms underlying prognosis and also assists patient stratification, treatment selection, and prediction of prognosis paths.

In the present paper, we applied a machine learning method, that has been previously shown to perform better than other state-of-the-art algorithms, on CM patients' vital status prediction based on RNA-Seq gene expression data. The accuracy achieved was 68.84% for the whole follow up period, significantly increasing with smaller time windows: 92.46% for 5-year vital status, 97.80% for 3-year and 100.00% for 1-year and significantly outperforming in all cases other implementation of Support Vector Machines and Random Forests.

Jiang et al [14] in their effort to predict survival time, thus treating the problem as a regression problem and not a classification one, achieved a C-statistic of 0.665 using the same data and 5-fold cross validation and similarly Sheng et al [22] achieved an AUC of 0.70, 0.69 and 0.68 for predicting 2, 3 and 5-year survival in their training set and Zeng et al [21] an AUC of 0.701 for 1 year, 0.726 for 3 years, and 0.745 for 5 years. We also plan to apply the regression version of the proposed method to get comparative results with these studies.

Another limitation of the proposed analysis is that for the calculation of the performance metrics, we have only relied on cross validation and have not used an external test set. Testing the generated prediction models in independent RNA-Seq gene expression data is also one of our immediate plans and the data collected in ebioMelDB will be useful for this task. Additionally, the gene signatures identified include 414, 13, 163 and 7 genes for the four prediction problems, respectively. We strongly believe that by experimenting with the goals (i.e increasing the significance of selected features minimization) of the employed machine learning method, we will manage to generate smaller gene signatures without having a classification performance drop.

Single omics may not be enough for estimating CM patient prognosis and an analysis based on multi-omics data or even better multi-modal data is necessary. This was essentially the vision before conceptualizing ebioMelDB. However, gene expression is the downstream product of other omics changes and “closest” to clinical outcomes as suggested in a number of studies and this is why it was chosen as our starting point.

Ultimately, integrating omics and clinical data with imaging data is promising and we aspire that ebioMelDB will assist the research community to make the limited relevant research wider.

Data availability is important when researchers want to apply machine learning methods for disease diagnosis, response to treatment and prognosis and for CM, clinical, biological and imaging data are scattered through the web. ebioMelDB is the first database to integrate the widest collections of RNA-Seq gene expression and clinical data with clinical and dermoscopy images, all manually curated and labelled with categories. In its future versions, ebioMelDB will host our under development predictive models in CM diagnosis, response to treatment and prognosis based on combinations of the different data types hosted.

Acknowledgements

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).

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

References

1. Rebecca, V. W., Somasundaram, R., & Herlyn, M. (2020). Pre-clinical modeling of cutaneous melanoma. *Nature communications*, *11*(1), 1-9.
2. Ali, A. R. H., Li, J., & Yang, G. (2020). Automating the ABCD Rule for Melanoma Detection: A Survey. *IEEE Access*, *8*, 83333-83346.
3. Barata, C., Celebi, M. E., & Marques, J. S. (2018). A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE journal of biomedical and health informatics*, *23*(3), 1096-1109.
4. Adegun, A., & Viriri, S. (2020). Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review*, 1-31.
5. Dimitriou, F., Krattinger, R., Ramelyte, E., Barysch, M. J., Micaletto, S., Dummer, R., & Goldinger, S. M. (2018). The world of melanoma: epidemiologic, genetic, and anatomic differences of melanoma across the globe. *Current oncology reports*, *20*(11), 1-9.
6. Xiong, J., Bing, Z., & Guo, S. (2019). Observed survival interval: a supplement to TCGA pan-cancer clinical data resource. *Cancers*, *11*(3), 280.
7. Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., ... & Kwong, L. N. (2015). Genomic classification of cutaneous melanoma. *Cell*, *161*(7), 1681-1696.

8. Papadodima, O., Kontogianni, G., Piroti, G., Maglogiannis, I., & Chatziioannou, A. (2019). Genomics of cutaneous melanoma: focus on next-generation sequencing approaches and bioinformatics. *Journal of Translational Genetics and Genomics*, *3*.
9. Bhalla, S., Kaur, H., Dhall, A., & Raghava, G. P. (2019). Prediction and analysis of skin cancer progression using genomics profiles of patients. *Scientific reports*, *9*(1), 1-16.
10. Lauss, M., Nsengimana, J., Staaf, J., Newton-Bishop, J., & Jonsson, G. (2016). Consensus of melanoma gene expression subtypes converges on biological entities. *Journal of Investigative Dermatology*, *136*(12), 2502-2505.
11. Pilla, L., Alberti, A., Di Mauro, P., Gemelli, M., Cogliati, V., Cazzaniga, M. E., ... & Maccalli, C. (2020). Molecular and Immune Biomarkers for Cutaneous Melanoma: Current Status and Future Prospects. *Cancers*, *12*(11), 3456.
12. Donnelly III, D., Aung, P. P., & Jour, G. (2019, December). The "-OMICS" facet of melanoma: heterogeneity of genomic, proteomic and metabolomic biomarkers. In *Seminars in cancer biology* (Vol. 59, pp. 165-174). Academic Press.
13. Zhang, L., Wang, Q., Wang, L., Xie, L., An, Y., Zhang, G., ... & Guo, X. (2020). OSskcm: an online survival analysis webserver for skin cutaneous melanoma based on 1085 transcriptomic profiles. *Cancer Cell International*, *20*, 1-8.
14. Jiang, Y., Shi, X., Zhao, Q., Krauthammer, M., Rothberg, B. E. G., & Ma, S. (2016). Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics*, *107*(6), 223-230.
15. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2014). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkol. 2015; 1A: A68–A77.
16. Chen, W., Cheng, P., Jiang, J., Ren, Y., Wu, D., & Xue, D. (2020). Epigenomic and genomic analysis of transcriptome modulation in skin cutaneous melanoma. *Aging (Albany NY)*, *12*(13), 12703.
17. Chai, H., Shi, X., Zhang, Q., Zhao, Q., Huang, Y., & Ma, S. (2017). Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genetic epidemiology*, *41*(8), 779-789.
18. Han, W., Huang, B., Zhao, X. Y., & Shen, G. L. (2020). Data mining of immune-related prognostic genes in metastatic melanoma microenvironment. *Bioscience reports*, *40*(11).
19. Meng, L., He, X., Zhang, X., Zhang, X., Wei, Y., Wu, B., ... & Xiao, Y. (2020). Predicting the clinical outcome of melanoma using an immune-related gene pairs signature. *PloS one*, *15*(10), e0240331.
20. Zhao, Y., Schaafsma, E., Gorlov, I. P., Hernando, E., Thomas, N. E., Shen, R., ... & Cheng, C. (2019). A leukocyte infiltration score defined by a gene signature predicts melanoma patient prognosis. *Molecular Cancer Research*, *17*(1), 109-119.
21. Zeng, Y., Zeng, Y., Yin, H., Chen, F., Wang, Q., Yu, X., & Zhou, Y. (2021). Exploration of the immune cell infiltration-related gene signature in the prognosis of melanoma. *Aging (Albany NY)*, *13*(3), 3459.
22. Sheng, Y., Tong, L., & Geyu, L. (2020). An immune risk score with potential implications in prognosis and immunotherapy of metastatic melanoma. *International Immunopharmacology*, *88*, 106921.
23. Garg, M., Couturier, D. L., Nsengimana, J., Fonseca, N. A., Wongchenko, M., Yan, Y., ... & Rabbie, R. (2021). Tumour gene expression signature in primary melanoma predicts long-term outcomes. *Nature communications*, *12*(1), 1-14.
24. Planell, N., Lagani, V., Sebastian-Leon, P., Van Der Kloet, F., Ewing, E., Karatha-nasis, N., ... & Gomez-Cabrero, D. (2021). STATegra: Multi-Omics Data Integration–A Conceptual Scheme With a Bioinformatics Pipeline. *Frontiers in Genetics*, *12*, 143.

25. Antonelli, L., Guarracino, M. R., Maddalena, L., & Sangiovanni, M. (2019). Integrating imaging and omics data: A review. *Biomedical Signal Processing and Control*, *52*, 264-280.
26. Peng, Y., Chu, Y., Chen, Z., Zhou, W., Wan, S., Xiao, Y., ... & Li, J. (2020). Combining texture features of whole slide images improves prognostic prediction of recurrence-free survival for cutaneous melanoma patients. *World journal of surgical oncology*, *18*(1), 1-8.
27. Maglogiannis, I., Kontogianni, G., Papadodima, O., Karanikas, H., Billiris, A., & Chatziioannou, A. (2021). An Integrated Platform for Skin Cancer Heterogenous and Multilayered Data Management. *Journal of Medical Systems*, *45*(1), 1-13.
28. Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2018). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, *23*(2), 538-546.
29. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5, 180161 (2018).
30. Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018, April). 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). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 168-172). IEEE.
31. Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., ... & Malvehy, J. (2019). BCN20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.
32. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., ... & Soyer, H. P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, *8*(1), 1-8.
33. Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M. F., & Petkov, N. (2015). MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, *42*(19), 6578-6585.
34. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013, July). PH 2-A dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5437-5440). IEEE.
35. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... & Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, *41*(D1), D991-D995.
36. Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S., & Chen, Y. (2008). GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, *24*(23), 2798-2800.
37. Gao, G. F., Parker, J. S., Reynolds, S. M., Silva, T. C., Wang, L. B., Zhou, W., ... & Noble, M. S. (2019). Before and after: comparison of legacy and harmonized TCGA genomic data commons' data. Cell systems, 9(1), 24-34.
38. Corthésy, J., Theofilatos, K., Mavroudi, S., Macron, C., Cominetti, O., Remlawi, M., ... & Dayon, L. (2018). An adaptive pipeline to maximize isobaric tagging data in large-scale MS-based proteomics. *Journal of proteome research*, *17*(6), 2165-2173.
39. Gudin, J., Mavroudi, S., Korfiati, A., Theofilatos, K., Dietze, D., & Hurwitz, P. (2020). Reducing opioid prescriptions by identifying responders on topical analgesic treatment using an individualized medicine and predictive analytics approach. *Journal of pain research*, *13*, 1255.