









# Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-Seq Datasets

Juan M. Gálvez , Daniel Castillo-Secilla , Luis J. Herrera , Olga Valenzuela , Octavio Caba , José C. Prados , Francisco M. Ortuño , and Ignacio Rojas 

**Abstract**—Many clinical studies have revealed the high biological similarities existing among different skin pathological states. These similarities create difficulties in the efficient diagnosis of skin cancer, and encourage to study and design new intelligent clinical decision support systems. In this sense, gene expression analysis can help find differentially expressed genes (DEGs) simultaneously discerning multiple skin pathological states in a single test. The integration of multiple heterogeneous transcriptomic datasets requires different pipeline stages to be properly designed: from suitable batch merging and efficient biomarker selection to automated classification assessment. This article presents a novel approach addressing all these technical issues, with the intention of providing new sights about skin cancer diagnosis. Although new future efforts will have to be made in the search for better biomarkers recognizing specific skin pathological states, our study found a panel of 8 highly relevant multiclass DEGs for discerning up to 10 skin pathological states: 2 healthy skin conditions a priori, 2 cataloged precancerous skin diseases and 6 cancerous skin states. Their power of diagnosis over new samples was widely tested by previously well-trained classification models. Robust performance metrics such as overall and mean multiclass F1-score outperformed recognition rates of 94% and 80%, respectively. Clinicians should give special attention to highlighted multiclass DEGs that have high gene expression changes present among them, and understand their biological relationship to different skin pathological states.

Manuscript received February 24, 2019; revised July 15, 2019 and October 11, 2019; accepted November 7, 2019. Date of publication December 23, 2019; date of current version July 2, 2020. The work of J. M. Gálvez was supported in part by the Government of Andalusia under Grant P12-TIC-2082 as part of the development of the research project “Advanced Computer Systems in Applications in the field of Biotechnology and Bioinformatics”, in collaboration with the Government of Spain under Grant RTI2018-101674-B-I00 and with the Feder Andalusia Operational Program Framework under Grant B-TIC-414-UGR18. (Francisco M. Ortuño and Ignacio Rojas are co-senior authors). (Corresponding author: Juan M. Gálvez.)

J. M. Gálvez, D. Castillo-Secilla, L. J. Herrera, and I. Rojas are with the Department of Computer Architecture and Computer Technology, University of Granada, 18071 Granada, Spain (e-mail: jmgg@ugr.es; cased@ugr.es; jherrera@ugr.es; irojas@ugr.es).

O. Valenzuela is with the Department of Applied Mathematics, University of Granada, 18071 Granada, Spain (e-mail: olgavc@ugr.es).

O. Caba and J. C. Prados are with the Institute of Biopathology and Regenerative Medicine (IBIMER), Center of Biomedical Research (CIBM), University of Granada, 18071 Granada, Spain (e-mail: ocaba@ugr.es; jcprados@ugr.es).

F. M. Ortuño is with the Clinical Bioinformatics Area, Fundación Progreso y Salud, Consejería de Salud, 41013 Seville, Spain (e-mail: franciscom.ortuno@juntadeandalucia.es).

Digital Object Identifier 10.1109/JBHI.2019.2953978

**Index Terms**—Gene expression, transcriptomic technologies, machine learning, feature selection, automated classification.

## I. INTRODUCTION

SKIN cancer is a worrying complex disease taking on a wide range of skin pathological states (SPSs). The complex heterogeneity of its occurrence is determined by the abnormal and out of control proliferation of specific cells (squamous, basal, Merkel, melanocyte, keratinocyte, etc.) that lead to the development of multiple skin cancerous pathologies. Among them, non-melanoma skin cancer (NMSC) related pathologies are the most frequent in order of incidences, pathologies in this group are led by basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and Merkel cell carcinoma (MCC) [1]. With regard to melanoma skin cancer (MSC), the main pathologies are primary melanoma (PRIMEL) and metastatic melanoma (METMEL), where METMEL has a higher mortality rate [2]. Recent epidemiological studies show a concerning global trend, the incidence and occurrence of both MSC and NMSC cases have already become the most common types of cancer in white populations [3]. This is supported by the statistical analysis of MSC rates on cohorts from United States whites, United Kingdom, Norway and Sweden which increased up to 3% annually during the last 3 decades [4]. With respect to NMSC cases, its incidence is around 20 times higher than MSC cases [5] despite being widely understudied. As a result of the fateful combination of both factors (incidence and occurrence), an extensive global alarm is being increased. Therefore, the possibility of suffering from any skin cancer type could be led by two main drivers. The first driver is the tumor evolution of other skin diseases previously considered precancerous states such as psoriasis (PS) [6], [7] or actinic keratosis (AK) [8]; the second driver is the tumor degeneration and mutation from healthy states such as normal skin (NSK) and nevus (NEV).

The narrow biological relationship among several SPSs may complicate the successful diagnosis of skin cancer. Certain researches have pointed out the difficulty in discerning among specific SPSs from the clinical, histological and molecular points of view: AK vs SCC [9], AK and SCC vs PRIMEL [10], SCC vs BCC and MSC [11], primary MCC (PMCC) vs metastatic MCC (MMCC) [12], etc. Different editions of the American Joint Committee on Cancer (AJCC) have gradually introduced the most outstanding clinical parameters for their diagnosis

(tumor mitotic rate, TNM classification, Breslow thickness, Clark levels, etc.). Consequently, the AJCC Cancer Staging Manual has been considered the gold standard by clinicians when making their diagnoses [13]. However, the way to diagnose this cancerous disease continues to be limited and each AJCC edition implies controversies and corrections on which are the best criteria to efficiently diagnose each SPS. Conversely, other studies insist on the possibility of differentiating them from the identification of gene expression patterns such as AK vs SCC [14]. Although previous studies show DEGs can discern among different pathologies, the biological complexity of the skin cancer may put its validity into question.

The opportunity to efficiently improve the discernment among multiple SPSs related to cancer from biological data involves taking into account a set of requirements. Firstly, different technological alternatives which allow gene expression to be quantified have to be inspected. Although microarray technology has been vastly used, RNA-seq technology is ultimately replacing it thanks to its various advantages described in the literature [15]. Nonetheless, the absence of open access datasets from experiments using the newest technologies suggests microarray analysis could still be considered. In addition to its low cost, it may not have been properly exploited yet by considering the combination of diverse skin cancer datasets containing samples of different SPSs. This fact gives the chance to reinforce the statistical robustness of the study as well as to obtain DEGs from a wider range of SPSs. This observation introduces the following challenge: how to adequately integrating data from both technologies in order to increase as much as possible the number of samples for each identified SPS of the study. Previous studies have underlined the good agreement among them in terms of similarity, complementarity and compatibility [16], [17]. In view of these advantages together with the proven consistency of applying multi-platform integration among both microarray platforms and technologies at gene expression level [18]–[22], this integrative approach is encouraged to continue carrying it out. However, the researchers have traditionally kept in mind the mandatory correction of eventual batch effects with the purpose of achieving an effective integration of multiple experiments over different microarray platforms [23], mainly coming from two manufacturers: Affymetrix [24] and Illumina [25]. By additionally taking into account experiments conducted on RNA-seq technology, the hypothetical influence of this factor may be modified in an unpredictable way. The treatment and the attempt of correction should never be disregarded; however, there is no certainty that a complete elimination of these effects will take place [26]. Among the multiple batch effect correction algorithms, *ComBat* [27] has proven to show the highest effectiveness when integrating microarrays [28] and, recently, has been validated in the integration of RNA-seq datasets from different sources: GTEx and TCGA projects [29]. In the case of favorably dealing with all these limitations, a new experimental challenge takes place: how to discern multiple SPSs by using changes in gene expression. Although hierarchical clustering highly helps in graphically showing such changes [30], methodological approaches based on multiclass classification are postulated as an innovative alternative when assessing the

validity of DEGs for simultaneously diagnosing multiple SPSs [31]. Finally, the use of feature selection algorithms must be explored with the objective of selecting only informative DEGs, that in many cases can dramatically reduce the search space.

Under the fulfillment of the previous requirements, the integration of microarray and RNA-seq technologies at gene expression level [32] opens new possibilities for skin cancer analysis. In particular, this advance could improve the understanding of the hypothetical biological relationships and differences among SPSs that may be discerned in a simple simultaneous analysis. Clinicians could directly benefit from its validity in multiple ways. Firstly, the suspicions about the patient tumor evolution from healthy skin states to cancerous states, even through pre-cancerous skin diseases, could be eventually assessed by presenting certain genetic susceptibility to change [33]. A patient-oriented medical service could be derived from the above by knowing the genetic signs. Consequently, unnecessary medications or medical treatments such as radiation therapies, excision surgeries or medications supply could be prevented [34]. Certainly, clinical diagnosis could be supported by an intelligent diagnosis tool that offers another complementary point of view [35], [36].

Although our novel methodological approach is thought to be applied on any multiclass problem, this work shows its validity by addressing the improvement of skin cancer diagnosis, thus taking into account all the requirements previously discussed and offering the benefits motivated above. The integration of different skin cancer datasets from microarray and RNA-seq technologies based on gene expression analysis has not been widely explored by the scientific community. First of all, an exhaustive sample search of multiple SPSs was carried out from public data repositories. Next, 22 microarray and 5 RNA-seq series containing 1090 samples in total were collected. However, after applying a strict quality control phase, only 968 samples passed and were subjected to the preprocessing phase: 666 samples from Affymetrix and Illumina microarray platforms and 302 samples from Illumina RNA-seq platforms. Subsequently, the sample integration considered only those genes sharing a common annotation for all the series selected for this study. After merging multiple batches and applying batch effect correction on them, the challenge was to efficiently find valid genes simultaneously discerning up to 10 SPSs: from a priori healthy states (NSK and NEV) to cutaneous carcinomas (BCC, ISCC, PMCC and MMCC) or melanomas (PRIMEL and METMEL), including skin diseases with a higher risk of tumor degeneration that have already been cataloged as precancerous states (AK and PS). From the assessment of a highly heterogeneous multiclass dataset of 968 samples and almost 7700 genes, a DEGs subset was identified by applying a novel one-vs-one (OVO) multiclass gene selection algorithm. This was achieved by means of consciously tuning critical and highly selective parameters. Specifically, log<sub>2</sub> fold change (LFC) and maximum number of selected DEGs (NMAX) among each pair of SPSs were considered. By relying on a widely used feature selection algorithm and assessing different subgroups of multiclass candidate DEGs, an ANOVA statistical test [37] assessed the influence of these critical parameters together with the use of different classification

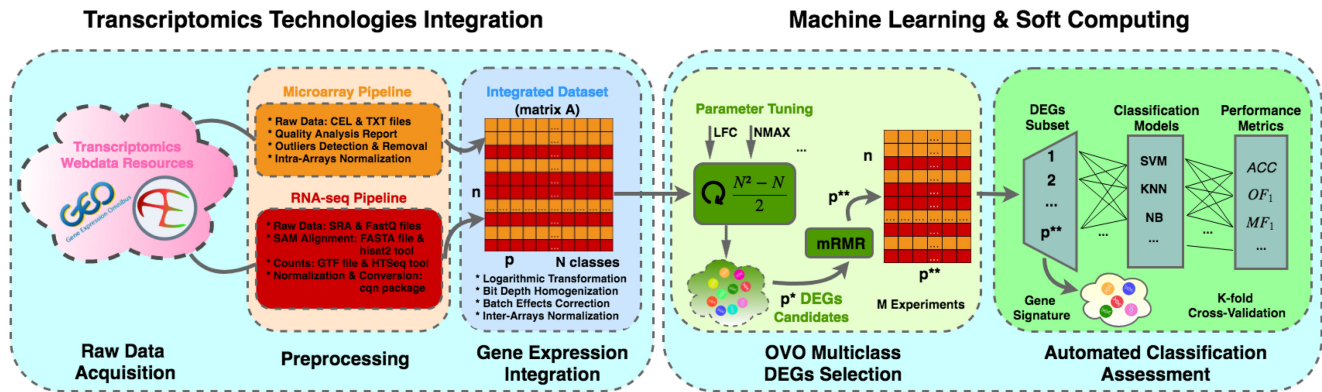


Fig. 1. Overall flowchart of the designed gene expression analysis pipeline. Two main bioinformatic tasks are addressed based on gene expression analysis: transcriptomics technologies integration and machine learning techniques application.

models and performance metrics. Finally, the biological relationship of these DEGs with skin cancer was determined by examining their functional properties and inspecting specific literature.

## II. METHODS

A flowchart of our approach is presented here (Fig. 1). Each of the experimental steps of this proposed pipeline will be subsequently addressed in the following subsections. For reproducibility purposes, this methodological approach can be run step-by-step by using different functionalities published under the KnowSeq R/Bioc package [38] together with several R scripts which have been included in the next repository: <https://github.com/jmgugr/skca-transcriptomics-integration/>

### A. Transcriptomics Technologies Integration

In order to obtain the integration of skin cancer datasets coming from different platforms and technologies, three steps have to be carried out (see left part in Fig. 1).

1) *Raw Data Acquisition*: One of the first steps involves carrying out an in-depth information search about skin cancerous pathologies and, subsequently, finding out the current availability of datasets. For example, AK and PS have been previously cataloged as precancerous skin diseases. Also, a wide range of SPSs related to cancer have been specified: from carcinomas (BCC, SCC or MCC) to melanomas (PRIMEL and METMEL), to even lymphomas or sarcomas. Next, the identification of transcriptomics webdata resources required to inspect the availability of the above SPSs together with healthy states (such as NSK or NEV) in public repositories such as NCBI GEO [39] and ArrayExpress [40] web platforms. Initially, samples on which drugs were applied, viruses were evaluated or were not directly extracted from tissue by means of punch biopsies or sliced sections were discarded. Moreover, only those SPSs for which a sufficiently representative number of samples were found and considered in order to increase the possibilities of characterizing their manifestation [41]. Under these considerations, Bowen's disease samples (also known as SCC in situ) were not finally considered (only two datasets containing data samples from this SPS were found, summing up to only 12 samples which

TABLE I  
TAXONOMIC CLASSIFICATION OF SKIN PATHOLOGICAL STATES FOR THE 968 COLLECTED RNA SAMPLES

Super-state	SPS	Microarray	RNA-seq	Integrated
Healthy state	NSK	151	84	235
	NEV	30	27	57
Non-melanoma skin cancer (NMSC)	BCC	43	0	43
	ISCC	69	14	83
	PMCC	26	0	26
	MMCC	23	0	23
Melanoma skin cancer (MSC)	PRIMEL	69	51	120
	METMEL	39	0	39
Precancerous skin disease	AK	29	23	52
	PS	187	103	290
Total		666	302	968

SPS = Skin pathological state, NSK = Normal skin, NEV = Nevus, BCC = Basal cell carcinoma, ISCC = Invasive squamous cell carcinoma, PMCC = Primary Merkel cell carcinoma, MMCC = Metastatic Merkel cell carcinoma, PRIMEL = Primary melanoma, METMEL = Metastatic melanoma, AK = Actinic keratosis, PS = Psoriasis.

was considered too low for the study). Finally, no representative number of lymphoma and sarcoma samples was found, so they were not considered in this study.

Since different microarray technologies and platforms were dealt with, several R packages from Bioconductor web platform [42] were used to acquire the RNA samples: *GEOquery* [43], *affy* [44] and *oligo* [45] for Affymetrix platforms and *lumi* [46] for Illumina platforms. In the case of RNA-seq series, SRA and FASTQ files containing raw information were directly downloaded in a programmatic manner before being preprocessed. Only those series whose samples were aligned to the GRCh37 reference genome, were considered for this study due to its greater current public availability. Specifically, the extensive RNA sample collection from 27 series used in this work led to the analysis of up to 10 SPSs (Table I).

Each of the series can be identified under accession ID, which shows most of them being submitted from United States and other countries where their population is predominantly white: Deutschland, Netherlands, Great Britain and Australia (Table II).

2) *Preprocessing*: This phase checks the quality of the samples under a detailed quality analysis process in order to remove potentially wrong samples. The quality of every microarray

TABLE II  
SERIES INFORMATION SELECTED FOR THIS STUDY FROM NCBI GEO AND ARRAYEXPRESS WEB PLATFORMS

Technology	Manufacturer	Series	Samples origin	Skin pathological states (Selected samples)	High quality samples	Excluded outliers		
Microarray	Affymetrix	GSE2503	Berlin (DEU)	ISCC (5), NSK (4), AK (3)	12	2		
		GSE3189	San Diego (USA)	NEV (16), NSK (6)	22	3		
		GSE6710	Berlin (DEU)	PS (13)	13	0		
		GSE7553	Tampa (USA)	BCC (15), PRIMEL (14), ISCC (11), NSK (4)	44	2		
		GSE13355	Ann Arbor (USA)	NSK (61), PS (56)	117	5		
		GSE14905	Gaithersburg (USA)	PS (31), NSK (16)	47	7		
		GSE15605	Nashville (USA)	PRIMEL (30), NSK (13), METMEL (2)	45	18		
		GSE30999	Spring House (USA)	PS (73)	73	12		
		GSE32407	New York (USA)	NSK (10)	10	0		
		GSE32924	New York (USA)	NSK (7)	7	1		
		GSE36150	Royal Oak (USA)	PMCC (5), MMCC (5)	10	5		
		GSE39612	Ann Arbor (USA)	PMCC (12), MMCC (5), ISCC (2), BCC (2)	21	15		
		GSE42109	New York (USA)	BCC (10)	10	1		
		GSE42677	New York (USA)	NSK (9), ISCC (5), AK (5)	19	1		
		GSE45216	London (GBR)	ISCC (27), AK (8)	35	5		
		GSE46517	Houston (USA)	METMEL (31), PRIMEL (25), NSK (6), NEV (6)	68	20		
		GSE50451	Bethesda (USA)	MMCC (13), PMCC (9)	22	1		
		GSE52471	New York (USA)	PS (14), NSK (10)	24	7		
		GSE53223	New York (USA)	NEV (8), NSK (5)	13	5		
		GSE82105	New York (USA)	METMEL (6)	6	0		
		RNA-seq	Illumina	GSE32628	Leiden (NLD)	ISCC (14), AK (13)	27	2
				GSE53462	Suwon (PRK)	BCC (16), ISCC (5)	21	5
				GSE54456	Ann Arbor (USA)	PS (89), NSK (80)	169	0
				GSE67785	Ann Arbor (USA)	PS (14)	14	0
				GSE84293	Houston (USA)	AK (10), ISCC (9)	19	0
				GSE98394	New York (USA)	PRIMEL (51), NEV (27)	78	0
				E-MTAB-5678	Brisbane (AUS)	AK (13), ISCC (5), NSK (4)	22	0
Integrated				968	122			

Samples purity for each skin pathological state was critically required and inspected. Manufacturer, technology and total number of samples/outliers are included.

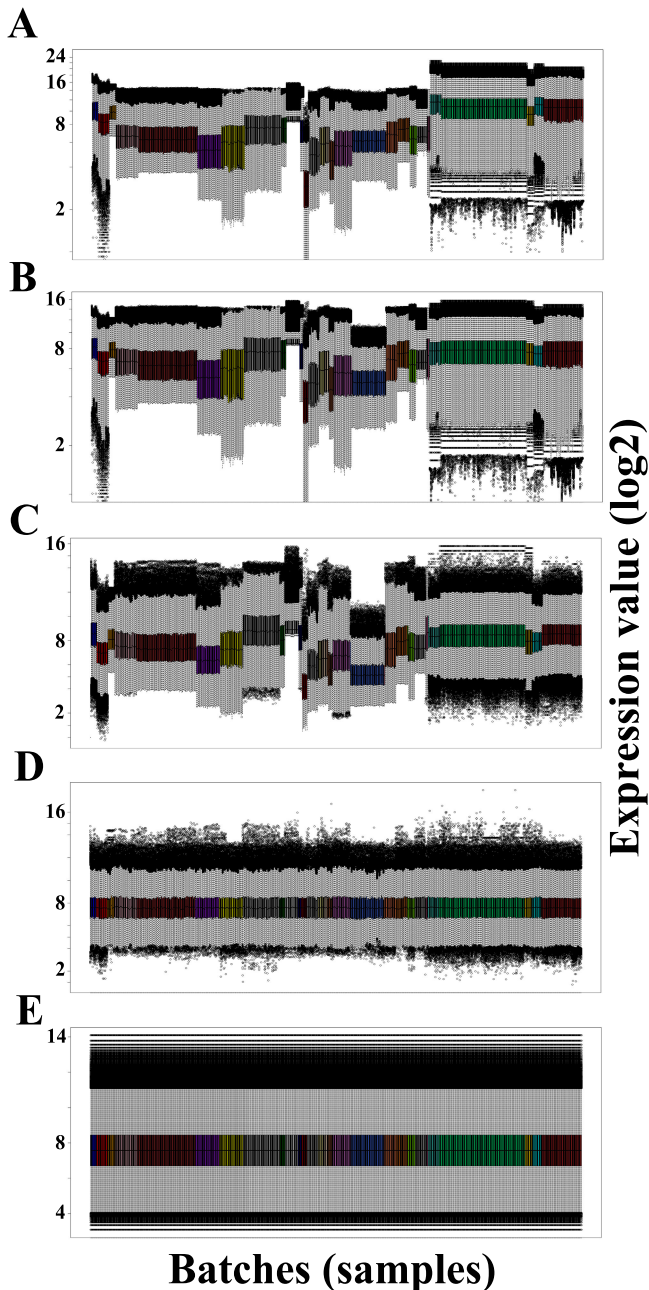
series was assessed using up to 6 quality tests: distance among samples, principal component analysis (PCA), Kolmogorov-Smirnov test based on the  $K_a$  parameter, density distribution plots, standard deviation of the samples intensities and Hoeffding's D-statistic (normally executed with  $D < 0.15$ ). These tests are available from the *arrayQualityMetrics* R package [47]. In order to discard all low quality samples (outliers), all of these tests were applied many times for each series. With respect to RNA-seq series, 5 samples were excluded by avoiding sample duplication. The total number of excluded samples from each series is specified in the last column of Table II.

Subsequently, each of the sequencing technologies requires a wide range of intra-array processing steps which have to be carefully performed when both are going to be finally integrated at gene expression level. Because of being processed from different platforms, a normalization procedure has to be applied on each microarray series. The Robust Multi-array Average (RMA) algorithm [48] was applied in this work by modularly performing background correction, normalization and summarization on the microarray data. The *rma* function from *affy* and *oligo* R packages was used for Affymetrix microarrays and the *lumiExpresso* from *lumi* R package was used for Illumina microarrays. Gene annotation of each series was provided by the *annotate* R package, which helps in mapping from the manufacturer chip identifiers to standardized symbols by using a wide range of annotation packages from the Bioconductor website. With respect to RNA-seq series processing, the proposed pipeline by Anders *et al.* [49] was partially followed, only changing certain tools. Once a large number of FASTQ and SRA files

are available, several tools such as *sra-toolkit* [50], *hisat2* [51], *bowtie2* [52], *samtools* [53] and *htseq* [54] were used to get read count files containing the located genes in each sample. Before obtaining these files, gene annotation was retrieved by means of *biomaRt* R package [55], a data-mining tool which allows connecting to the *Ensembl* database [56]. After all these steps, other R packages such as *cqn* [57] helped in correcting and normalizing GC content bias, and *NOISeq* [58] was used to calculate the gene expression values.

3) *Gene Expression Integration*: After preprocessing each of the microarray and RNA-seq series individually, additional requirements have to be considered before inter-array normalization and integration [59]. First of all, each of the expression values of the genes transcribing the same gene identifier have to be summarized in a single value. In order to be consistent in assessing the impact of each gene selected in our analysis, all transcripts were gathered by applying the mean of them to each series separately. This parameter was selected after performing a comparative study versus median as it was done in our previous work [31], showing no statistically significant differences in classification performance. Next, several simultaneous steps were carried out on the 27 series (Fig. 2). 28 batches were established because different samples from GSE42677 series were processed by two different platforms.

Firstly, logarithmic transformation was performed on 2 series (GSE2503 and GSE3189) in order to adjust the scale of the gene expression values, establishing base 2 for all the batches (Fig. 2A). Next, 16-bit depth homogenization was applied (Fig. 2B) after previously analyzing the maximum



**Fig. 2.** Series processing procedure for gene expression integration: (A) logarithmic transformation, (B) 16-bit depth homogenization, (C) complete cases selection along the batches, (D) batch effect correction with *ComBat* and (E) inter-array normalization with *normalizeBetweenArrays*.

value of gene expression for each series in function of the platform, establishing different consensus values in the bit depth: 20-bit depth for Human Genome U133A Array platform (GSE2503, GSE3189, GSE6710 and GSE46517), 16-bit depth for Human Genome U133 Plus 2.0 Array platform (GSE7553, GSE13355, GSE14905, GSE15605, GSE30999, GSE32924, GSE39612, GSE42677, GSE45216, GSE50451, GSE53223 and GSE82105), 16-bit depth for Human Genome U133A 2.0 Array platform (GSE32407, GSE42109, GSE42677

and GSE52471), 12-bit depth for Human Exon 1.0 ST Array platform (GSE36150), 16-bit depth for HumanAll platform (GSE32628 and GSE53462), 22-bit depth for Genome Analyzer platform (GSE54456), 20-bit depth for Genome Analyzer IIX platform (GSE67785), 24-bit depth for HiSeq 2000 platform (GSE84293 and E-MTAB-5678) and 22-bit depth for HiSeq 2500 platform (GSE98394). Thereafter, by having previously established a common gene annotation for all the considered series, only common genes from all the samples coming from the series / batches were identified and selected. At this point, batch effect correction should be considered because hypothetical batch effects could be appearing among all 28 batches considered (Fig. 2C). In order to deal with this issue, *ComBat* method [27] from *sva* R package [60] was considered, correcting and establishing a consistent sample distribution along all the samples from all batches (Fig. 2D). Finally, an inter-array normalization was applied by means of *normalizeBetweenArrays* function from *limma* R package [61]. This achieves consistency among all the samples put together and forces an identical empirical distribution on each of them based on quantile normalization (Fig. 2E). Before any new sample is properly assessed by this procedure, all these transformations are completely necessary and have to be applied in the same way. At the end of this procedure, the whole integrated dataset formed by  $p$  common genes and all  $n$  quality samples selected among  $N$  classes is achieved (matrix A in Fig. 1).

## B. Machine Learning and Soft Computing

Bioinformatics researches and recent biological problems have been successfully benefited from the use of machine learning and soft computing techniques [62] in a wide range of topics such as expression profiling identification, feature selection and classification [63], protein sequences [64] and DNA sequences [65]. As the number of biological experiments and applications using high-throughput technologies continues to increase, this type of techniques for knowledge discovery will find new applications [66], [67].

1) *OVO Multiclass DEG Selection*: Traditionally, the gene selection from expression profiles analysis deals with the curse of dimensionality problem ( $np$ -hard), as it pits few  $n$  samples against thousands of  $p$  genes [68]. This issue becomes even more challenging when increasing the number of SPSs (in our work,  $N$ ) (see nomenclature in Fig. 1).

For the purpose of addressing such challenge, this work presents a novel, simple and intuitive one-vs-one (OVO) multiclass DEGs selection approach based on the assessment of all possible pair comparisons of SPSs. This work defines the comparison of two SPSs as class pair comparison (CPC). The criterion for selecting DEGs for each CPC is a high LFC, which means higher discernment power at the gene expression level. The DEGs selection process ensures this criterion is satisfied by tuning the two parameters *LFC* and *NMAX*. On the one hand, *LFC* establishes a minimum threshold value for genes to be considered as DEGs throughout all CPCs. On the other hand, *NMAX* indicates the maximum number of DEGs selected for each CPC. An additional threshold can be established using

*p*-value (PV). However, a constant value of 0.001 was set in the experimental setup of this work. By extending to a problem of  $N$  SPSs, the total number of CPCs amounts to  $(N^2 - N)/2$ . In particular, this work simultaneously analyzes 10 SPSs, which gives 45 CPCs. The expected maximum number of DEGs, globally for all the CPCs, would sum up to  $NMAX * (N^2 - N)/2$ . This novel multiclass OVO approach can be seen as a step forward with respect to the classical gene selection process, which is exclusively controlled by *PV* and *LFC*. The lack of capacity of the selected DEGs to discern among specific CPCs or different SPS subsets can be easily avoided by tuning the new proposed *NMAX* parameter. The union of all the DEG sets after considering each CPC needs to take into consideration the identification of repeated DEGs. These can occur as some genes may present a strong difference of gene expression for several CPCs. Nevertheless, such DEGs coincidences would help in reducing even further (up to  $p^*$ ) the final candidate multiclass DEGs (where  $p^* \leq NMAX * (N^2 - N)/2 \ll p$ ).

In order to strengthen the selection of DEGs, a number of experiments ( $M = 10$  in this work) were performed, splitting in each of them the whole integrated dataset into two datasets: 90% for training and validation, and the remaining 10% for testing, in a cross-validation manner. Similar representativeness of each SPS was ensured within each of the dataset folds. The feature selection and parameter tuning processes were initially applied on the training dataset for each of these  $M$  experiments, thus returning different DEGs sets for each LFC and *NMAX* combination. With the aim of improving the reliability and the interpretability of the subsequent results, only those  $p^*$  common genes matching all the  $M$  experiments for each parameter combination were selected. This choice discards spurious DEGs only emerging in specific executions and prevents subsequent classification biases. Before evaluating the different  $p^*$  common gene set (finally naming  $p^*$  to this set) within each of the  $M$  experiments, an additional assessment of their informative power was performed by means of the minimum-Redundancy Maximum-Relevance (mRMR) feature selection algorithm [63]. This algorithm returns a ranking according to the criterion of placing first those DEGs with the most relevant and the lowest redundant information among themselves with respect to the class variable. Then, from this ranking, proper assessment, described in the next subsection, allowed selecting a total of  $p^{**}$  genes from the previous set of  $p^*$  genes.

In summary, twofold DEG selections were carried out: firstly, reducing the computational complexity from  $p$  thousands of genes to the  $p^*$  most reliable candidate DEGs of the disease; secondly, exclusively selecting those  $p^{**}$  DEGs with higher informative capability for the intelligent diagnosis (see right part in Fig. 1).

2) **Automated Classification Assessment:** Three classification techniques assessed the informative power of different DEGs subsets from the ranking returned by mRMR: Support Vector Machines (SVM) [69], K-Nearest Neighbour (KNN) [70] and Naive Bayes (NB) [71]. K-fold cross validation technique (K-fold CV, where  $K = 10$ ) [72] was used on the training set of each  $M$  experiment with the purpose of providing a realistic performance of the DEGs on new unseen data. Optimal

hyperparameters were calculated for these methodologies:  $\sigma$  (kernel width) and  $\gamma$  for SVM, and  $k$  for KNN. The 10-fold CV classification assessment was repeated 10 times by randomly shuffling the dataset. The advantages of this process are twofold: first, this is statistically robust because of asymptotic convergence to a reliable estimation of the classifier performance [73]; and second, this prevents from achieving overfitting when assessing training and testing data. Finally, three metrics were used in order to measure the recognition rate by combining each classifier in association with different DEGs set sizes: accuracy (*ACC*), overall F1-score (*OF<sub>1</sub>*) and mean multiclass F1-score (*MF<sub>1</sub>*). These are calculated by using Equation (1), (2) and (3), respectively. Each of these metrics can be expressed as a function of certain parameters (precision ( $P$ ) and recall ( $R$ )) or different rates ( $T_p$ ,  $T_n$ ,  $F_p$  and  $F_n$ ) which can be identified from a confusion matrix of  $N$  classes:

$$ACC = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (1)$$

$$OF_1 = \frac{2 * P * R}{P + R} = \frac{2 * T_p}{2 * T_p + F_p + F_n} \quad (2)$$

$$MF_1 = \frac{\sum_{i=1}^N F_1^{class}(i)}{N} \quad (3)$$

The metrics related to  $F_1$ -score [74] were considered particularly suited and robust for the multiclass study tackled, as they provide a better measurement of the recognition rate of each of the classes under unbalanced data. With regard to dealing with it, data balancing techniques such as SMOTE [75] were considered. However, no significant performance improvement was achieved, leading to be discarded and avoiding the introduction of additional complexity and artificial data.

Next, in order to assess the influence of the multiple factors considered for identifying multiclass DEGs, an ANOVA statistical test was performed over the entire dataset. Although factors such as assessed dataset type (TYPE), analyzed K-fold cross validation (KFOLD) or  $M$  experiment performed (EXPERIMENT) were also evaluated by this test, 4 factors were specifically highlighted because of their further relevance in the subsequent analysis. On the one hand, *LFC* and *NMAX* parameters were subjected to evaluation by tuning the proposed algorithm. On the other hand, the hypothetical differences of applying different classifiers in combination with a number of DEGs set sizes (*GenMax*) were also inspected by means of this test. By checking the validity of each factor (*LFC*, *NMAX*, classifier and *GenMax*), the different performance metrics (*ACC*, *OF<sub>1</sub>* and *MF<sub>1</sub>*) were measured for both training and test sets.

Finally, a functional enrichment analysis was performed by means of DAVID 6.8 [76] in order to functionally annotate and interrelate the obtained DEGs using Gene Ontology (GO) terms [77].

### III. RESULTS AND DISCUSSION

By taking into account the integration at gene expression level from 22 microarrays and 5 RNA-seq series containing multiple

**TABLE III**  
ANOVA STATISTICAL TEST FOR MF1 PERFORMANCE METRIC

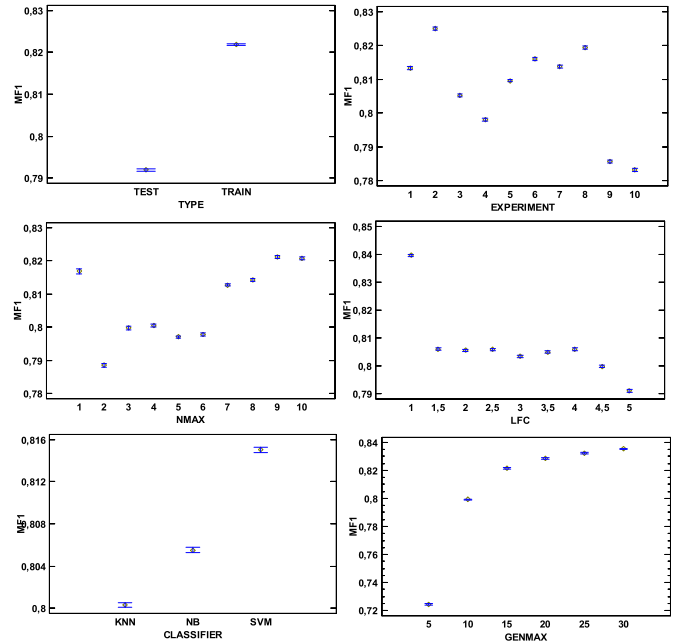
Source (Main Effects)	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
<b>A: TYPE</b>	61.4457	1	61.4457	22681.85	<b>0.0000</b>
<b>B: EXPERIMENT</b>	48.0765	9	5.34183	1971.87	<b>0.0000</b>
C: KFOLD	0.00597	9	0.00066	0.24	0.9878
<b>D: CLASSIFIER</b>	10.1689	2	5.08447	1876.86	<b>0.0000</b>
<b>E: LFC</b>	42.0707	8	5.25884	1941.23	<b>0.0000</b>
<b>F: NMAX</b>	30.6972	9	3.4108	1259.05	<b>0.0000</b>
<b>G: GENMAX</b>	419.897	5	83.9794	30999.84	<b>0.0000</b>
RESIDUAL	739.442	272955	0.00270		
TOTAL (CORRECTED)	1398.59	272998			

Type III sums of squares was chosen and the contribution of each factor was measured having removed the effects of all other factors. P-values tested the statistical significance of each of the factors. Since 6 P-values are less than 0.05, these have a statistically significant effect on MF1 at the 95.0% confidence level (in bold). F-Ratios are based on residual mean square error.

SPSs related to cancer, the opportunity to determine the skin cancer gene signature of up to  $N = 10$  SPSs, formed by highly reliable multiclass DEGs, has been addressed in this work. The experimental analysis of this study has been conducted under the proposal of a novel OVO multiclass DEGs selection algorithm, which has been thoroughly tested by means of a complete and powerful ANOVA statistical test. The interpretation of the results obtained from this analysis has been used to select suitable parameter settings. By tuning our proposed algorithm, this study was focused on assessing the informative power of the  $p^*$  identified multiclass DEGs. After selecting  $p^{**}$  multiclass DEGs from the previous one, their biological relationship to skin cancer was finally determined. This discussion has been guided on presenting all the results derived from the procedure above.

### A. Impact of Tuning Algorithm Parameters

The statistical significance of each considered and highlighted factor (NMAX, LFC, GenMax, Classifier) was confirmed by means of the ANOVA statistical test, showing the influence of each of them on the classification performance (Table III). The most significant differences were exclusively explained by checking the scale depth when using MF1 (Fig. 3). While the lowest NMAX parameter value reflected one of the highest classification performances and discarded the consideration of a wide range of DEGs for each of the 45 CPCs, the impact of tuning LFC helped to elucidate the disadvantage of selecting high threshold values because the MF1 value dropped by more than 3%. Classification models results ranged from 80% to 82%, establishing these performances around 10 genes (see Classifier and GenMax factors in Fig. 3). Similar statistical results and distribution for each factor were achieved for ACC and OF1 as well, and these can be facilitated under petition. Next, in order to present the utility of the proposed algorithm in this work, a choice of parameters was required. The decision was motivated under the criterion of restrictively selecting DEGs while preserving the information of all considered SPSs for this study. For this purpose,  $NMAX = 1$  was established as it presented one of the highest recognition rate for each performance



**Fig. 3.** ANOVA statistical test results for MF1 in function of different factors: Type, Experiment, NMAX, LFC, Classifier and GenMax. All these factors were determined as significant statistically.

metric assessed (as clearly showed and supported by the results of ANOVA statistical test), leading to drastically reducing the computational complexity to a maximum of  $(N^2 - N)/2 = 45$  highly discerning DEGs. This choice prevents of arbitrarily tuning LFC and relying decision power on it in search of a sufficient threshold for discerning among multiple SPSs. Furthermore, this decision may avoid the removal of DEGs to discern those hardly distinguishable CPCs when applying highly restrictive LFC values. Hereafter, these setting parameters were used to identify the candidate multiclass DEGs and present a potential gene signature of skin cancer.

### B. Selection of Informative DEGs

Although up to 45 genes could have potentially been returned by our proposed algorithm under the selected configuration, exclusively  $p^* = 10$  candidate multiclass DEGs appeared as common genes from the intersection of DEGs for each of the  $M = 10$  experiments performed, as many of these genes were highly discriminating among several CPCs. However, in order to reduce the repertoire of candidate DEGs set for intelligent diagnosis, the informative capability of different subgroups of up to  $p^*$  DEGs ordered by means of mRMR, was subjected to an automated classification assessment. This algorithm then established the following DEGs ranking: *MLANA*, *LTF*, *MMP1*, *ADAMTS3*, *LY6D*, *SCGB2A2*, *KRT14*, *PI3*, *PMEL* and *S100A7*. As a result, the classification results are presented when increasing the size of DEGs set following the previously established ranking, showing asymptotic convergence for the different performance assessments (Fig. 4). Differences among classifiers are not appreciated given the high discernment quality provided by the selected DEGs. By reducing the complexity of the study, the

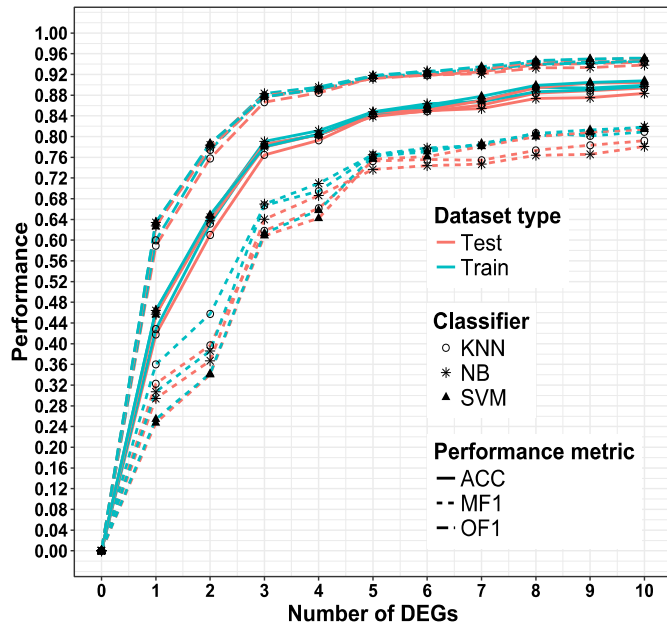


Fig. 4. Evolution of the recognition rate for training and test datasets. Three classification models (SVM, KNN, and NB) were assessed by means of several performance metrics (ACC, OF1, and MF1) when considering different subgroups of DEGs ranked by mRMR algorithm.

subsequent experimental analysis was limited to considering the first  $p^{**} = 8$  DEGs given that the further average improvement of MF1 per gene is lower than 0.6%. The results associated with this size of DEGs set even improved those showed by *GenMax* parameter for ANOVA test, outperforming recognition rates of 94% OF<sub>1</sub> and 80% MF<sub>1</sub> when considering any classifier.

Afterwards, with the purpose of knowing the overall discernment capabilities of the 8 multiclass candidate DEGs, the number of SPSs and CPC cases being covered by each one of them and the information of the highest |LFC| for any CPC was summarized (Table IV).

### C. Recognition of SPSs

Despite establishing parameter settings which help in exposing DEGs to discern from each CPC, difficulties in distinguishing among certain SPSs still can not be avoided. Most CPCs can be properly discerned from any of the 8 DEGs, presenting significant LFC values (Fig. 5). However, there is a small set of CPCs which are harder to distinguish when examining changes at gene expression level such as ISCC vs AK (LFC < 2) or PMCC vs MMCC (LFC < 1). This occurs when trying to offer a reliable diagnosis among a lot of SPSs which are close at the biological level.

By extensively checking how a new unseen sample could be classified, the different classification models assessed the 8 highlighted DEGs set (Fig. 6). The recognition rates confirm the real challenge of properly discerning the CPC cases previously highlighted, although presenting accuracy differences among models when classifying certain SPSs (for example, ISCC achieves 72% for NB, 76% for KNN and 77% for SVM). On the one hand, 3 SPSs achieved high recognition rates for SVM classification model: NSK (97%), BCC (~100%) and PS

TABLE IV  
STATISTICAL SUMMARY OF THE FIRST 8 MULTICLASS DEGS RETURNED BY MRMR ALGORITHM

Gene Symbol	SPSs	CPCs (%)	$ \mu_{LFC} \pm \sigma_{LFC} $	$[PV_{MIN}, PV_{MAX}]$
MLANA	7	8 (17.8)	$4.86 \pm 0.97$	[1.30E-157, 8.52E-86]
LTF	5	4 (8.9)	$4.81 \pm 0.29$	[3.87E-186, 2.70E-94]
MMP1	7	7 (15.6)	$4.67 \pm 1.48$	[1.34E-77, 8.30E-14]
ADAMTS3	4	3 (6.7)	$4.02 \pm 0.27$	[3.91E-230, 1.67E-160]
LY6D	5	5 (11.1)	$5.35 \pm 0.33$	[4.55E-135, 5.71E-121]
SCGB2A2	6	5 (11.1)	$5.48 \pm 0.69$	[7.58E-121, 4.41E-89]
KRT14	7	6 (13.3)	$6.24 \pm 0.56$	[2.42E-264, 1.96E-189]
PI3	7	6 (13.3)	$6.39 \pm 1.07$	[9.99E-220, 7.23E-73]

Average and standard deviation values for LFC parameter ( $|\mu_{LFC} \pm \sigma_{LFC}|$ ) as well as minimum and maximum values for PV parameter ( $[PV_{MIN}, PV_{MAX}]$ ) were also included when the additional statistical restriction for our approach ( $PV \leq 0.001$ ) was fulfilled. The number of SPSs and CPC cases being covered for each DEG together with associated statistical parameters are included. Only 1 CPC was not covered by presenting any DEG with significant statistical values: PMCC vs MMCC. SPS = skin pathological state, CPC = class pair comparison, PV = P-Value, PMCC = primary Merkel cell carcinoma, MMCC = metastatic Merkel cell carcinoma.

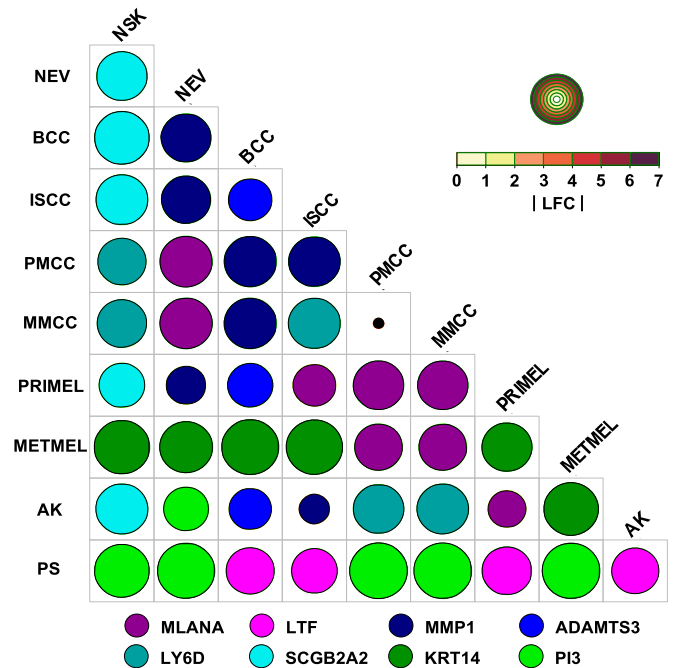


Fig. 5. Distribution map of the 8 multiclass DEGs set. Highest |LFC| value for each CPC by considering NMAX = 1 and applying mRMR algorithm. Circle size and color are correlated with |LFC| value and multiclass DEG with highest |LFC|, respectively. CPC, class pair comparison.

(~98%). On the other hand, recognition rates dropped for the 7 remaining SPSs mainly due to the confusion with another SPS as previous studies had already advanced [9]–[12]: NEV (~83% and confused with NSK above 4%), ISCC (77% and confused with AK above 20%), PMCC (~58% and confused with MMCC above 37%), MMCC (45% and confused with PMCC above 54%), PRIMEL (91% and confused with METMEL above 2%),



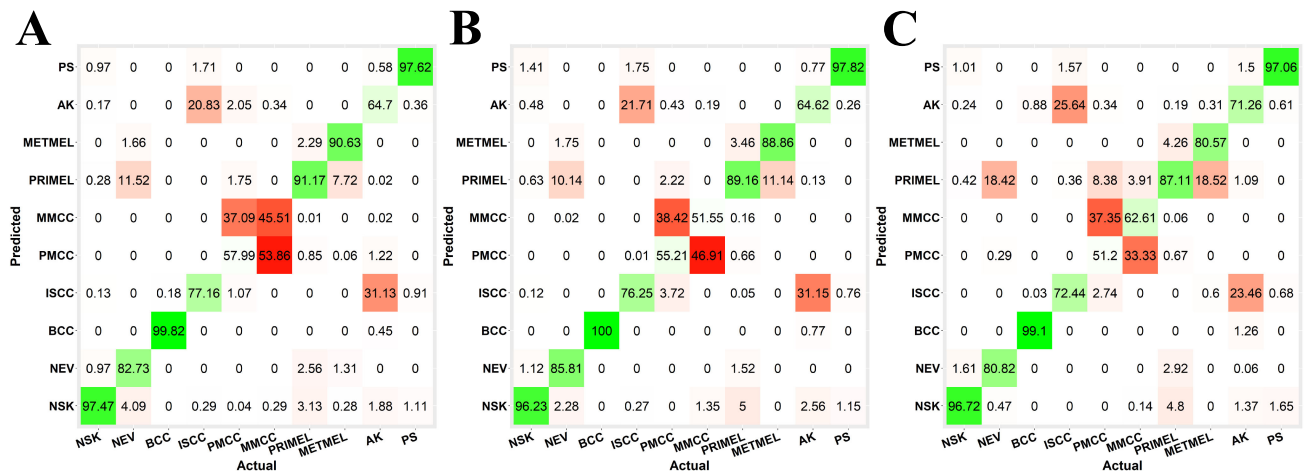


Fig. 6. Different classification models were assessed: (A) SVM, (B) KNN and (C) NB. For each designed model, the 8 highlighted multiclass DEGs set were selected and assessed by 10-fold CV for discerning 10 SPSs. SVM = Support Vector Machines, KNN = K-Nearest Neighbors, NB = Naive Bayes, CV = Cross-Validation, SPS = Skin pathological state.

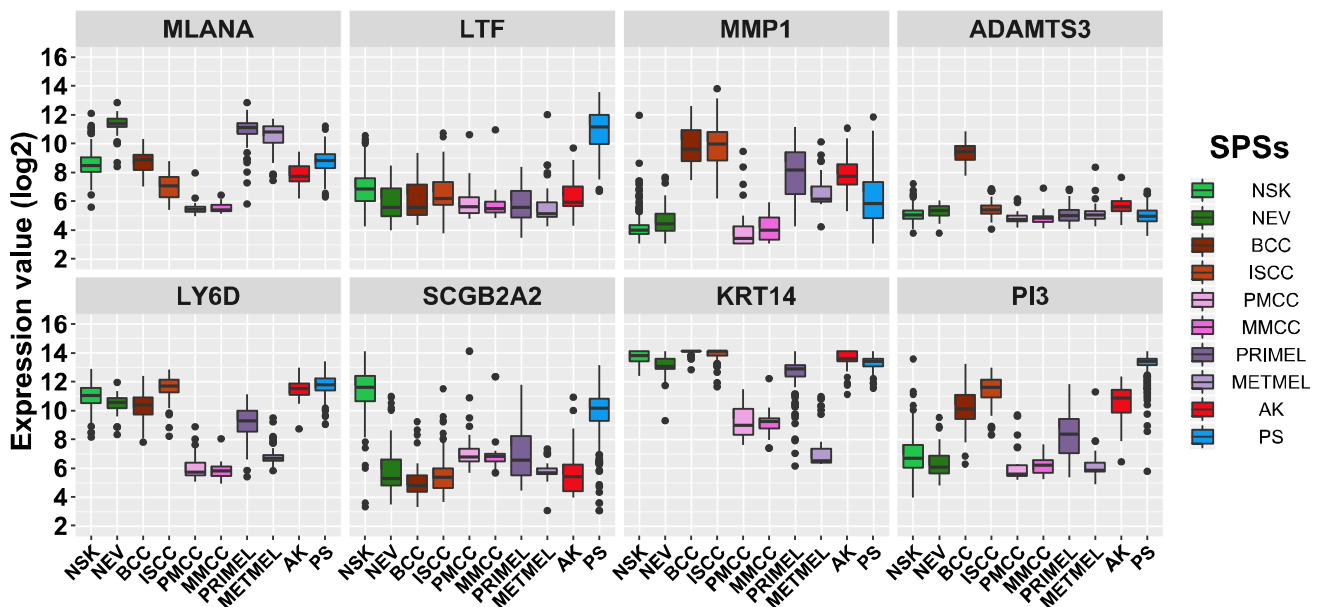


Fig. 7. Expression level of the 8 multiclass DEGs set. Highlighted DEGs by our approach are ordered from left to right and from top to bottom by the ranking returned by mRMR. SPS = Skin pathological state.

METMEL (90% and confused with PRIMEL above 7%) and AK (~65% and confused with ISCC above 31%). This fact remarks the difficulty of achieving reliable DEGs between precancerous and invasive states as they present molecular similarities. By considering the fusion of certain CPCs (for example, MCC formed by PMCC and MMCC, MSC formed by PRIMEL and METMEL or combining ISCC and AK), the recognition rates would have considerably increased the percentage up to 87–99% for these skin super-states in a more generalized study.

#### D. Determination of Potential Target Genes

One of the main justifications for separating specific SPSs is by using relevant biomarkers of their occurrence from gene expression analysis. In this case, our approach highlighted the informative capacity of these 8 candidate multiclass DEGs for an overall diagnosis of suffering from skin cancer (Fig. 7).

In view of these results, certain multiclass DEGs such as MLANA, MMP1, LY6D or PI3 appeared down-expressed for both SPSs and, among others, may discern better PMCC and MMCC with respect to other SPSs (Fig. 5). All these genes have previously proven to be of great importance for expression pattern characterization and skin cancer diagnosis: from inhibition in SCC (MLANA), positive dysregulation in BCC and AK (MMP1) to correlated overexpression in SCC and PS (PI3) [78]–[80]. Therefore, a preventive clinical analysis of these genes could help to avoid erroneous therapies by examining their hypothetical involvement in other SPSs addressed by this study.

#### E. Biological Interpretation of the Multiclass DEGs

In order to understand the functional properties of the 8 highlighted DEGs, an enrichment analysis based on GO terms

**TABLE V**  
FUNCTIONAL ENRICHMENT ANALYSIS FOR THE 8 MULTICLASS DEGs USING GO TERMS

Ontology	GO ID	GO Term	# Genes (%)	PV	Gene Symbol
BP	GO:0006508	Proteolysis	4 (50.0)	8.75E-3	LTF, MMP1, ADAMTS3, PI3
	GO:0030574	Collagen catabolic process	2 (25.0)	1.92E-2	ADAMTS3, MMP1
	GO:0032963	Collagen metabolic process	2 (25.0)	3.23E-2	ADAMTS3, MMP1
	GO:0044236	Multicellular organism metabolic process	2 (25.0)	3.90E-2	ADAMTS3, MMP1
	GO:0044243	Multicellular organism catabolic process	2 (25.0)	2.13E-2	ADAMTS3, MMP1
	GO:0044259	Multicellular organismal macromolecule metabolic process	2 (25.0)	3.38E-2	ADAMTS3, MMP1
CC	GO:0005576	Extracellular region part	5 (62.5)	3.95E-2	LTF, MMP1, ADAMTS3, KRT14, PI3
	GO:0005578	Proteinaceous extracellular matrix	3 (37.5)	5.57E-3	MMP1, ADAMTS3, PI3
	GO:0031012	Extracellular matrix	3 (37.5)	1.17E-2	MMP1, ADAMTS3, PI3
	GO:0031982	Vesicle	5 (62.5)	1.89E-2	MLANA, LTF, ADAMTS3, KRT14, PI3
	GO:0031988	Membrane-bounded vesicle	5 (62.5)	1.64E-2	MLANA, LTF, ADAMTS3, KRT14, PI3
	GO:0044421	Extracellular region part	5 (62.5)	2.12E-2	LTF, MMP1, ADAMTS3, KRT14, PI3
MF	GO:0004175	Endopeptidase activity	3 (37.5)	1.19E-2	LTF, MMP1, ADAMTS3
	GO:0004222	Metalloendopeptidase activity	2 (25.0)	3.95E-2	MMP1, ADAMTS3
	GO:0008233	Peptidase activity, acting on L-amino acid peptides	3 (37.5)	2.51E-2	LTF, MMP1, ADAMTS3
	GO:0070011	Peptidase activity, acting on L-amino acid peptides	3 (37.5)	2.35E-2	LTF, MMP1, ADAMTS3

Fisher's exact statistical test was performed to determine their significance ( $PV < 5E-2$ ). GO = Gene Ontology, PV = P-Value, BP = biological process, CC = cellular component, MF = molecular function.

was performed from DAVID Bioinformatics Database [75]. The three GO ontologies for biological processes (BP), cellular components (CC) and molecular functions (MF) were considered for our analysis. A total of 6 BPs, 6 CCs and 4 MFs were determined to be significant throughout these genes (Table V). As shown, MMP1, ADAMTS3 and LTF genes are highly related in terms of their proteolysis process and endo- and metalloendo-peptidase activity. According to the activity of proteolytic enzymes, this occurrence has been associated with angiogenesis and tumor progression of skin cancer [80], [81]. Subsequently, by exhaustively inspecting specific literature, the biological relationship of the 8 highlighted DEGs with skin cancer was consulted. On the one hand, the most remarkable inquiries underlined the dysregulation of up to 6 DEGs in MSC cases [83], [84] and development risk [85] (except ADAMTS3 and PI3) and the implication of up to 5 DEGs in PS development or inflammatory processes [80], [86] (except MLANA, ADAMTS3 and KRT14). On the other hand, the differentiating role of specific DEGs in NMSC cases was highlighted: the overexpression of ADAMTS3 in BCC [87] or the hypothetical implication of KRT14 in the malignant transformation of potential stem cells as origin of MCC [88]. Based on all these precedent evidences and the results shown (Fig. 7), the 8 multiclass DEGs highlighted by this approach should be particularly taken into account by being related to tumorigenesis and pathogenesis of skin cancer. Concretely, MLANA has been remarkably demonstrated to be upregulated in NEV [83], inhibited in SCC [78] and differentiated between MCC and PRIMEL by highlighting absence and overexpression by means of immunohistochemical analysis [89]. Further, multiple genetic dysregulations of DEGs have been reported in several studies: from the downregulation of LY6D, SCGB2A2 and KRT14 in METMEL with respect to PRIMEL [82] to the dysregulation in SCC versus NSK by showing inhibition of MLANA and SCGB2A2 or overexpression of MMP1 and PI3 [78], [90], [91]. Finally, the dysregulation of certain DEGs has been interestingly reflected in both SCC and PS in a similar way: from inhibition of SCGB2A2 together with overexpression of

MMP1 and PI3 [80] to slight and strong upregulation of LTF in SCC and PS, respectively [80], [86]. Because of being a chronic inflammatory skin disease, special attention should be paid to the psoriasis evolution because the cancer development also generates inflammatory reactions around surrounding tissue [7]. From the preventive point of view, clinicians should remain attentive to the high gene expression variability of these specific DEGs by observing changes between NSK, PS and diverse SPSs related to cancer (see gene expression changes for all these DEGs in Fig. 7). In accordance with our results, this multiclass DEGs subset could represent a genetic signature offering clues about the overall state of the disease.

#### F. Limitations of the Approach

Although the validity of gene expression analysis has been proven by an enormous amount of scientific publications, certain limitations may be remarked when applying multiclass classification and transcriptomics resources integration. On the one hand, the identification of relevant multiclass DEGs is restricted to the establishment of proper thresholds for the discernment among CPCs by using statistical conditions such as LFC or PV. However, it is completely necessary to inspect the gene expression levels in order to avoid the selection of DEGs with similar levels and avoid mistakes under classification assessment. This concern disappears when final clinical diagnosis is doubtful among exclusively two specific SPSs (for example, PMCC vs MMCC, etc.), in which case gene expression analysis may find highly discerning DEGs making them outstanding target genes for therapeutic treatments. On the other hand, the consideration of multiple heterogeneous sources of transcriptomics data may lead to removal of relevant DEGs after their integration, due to the specific platform requirements. This eventuality could be widely assessed at the expense of discarding biological samples as long as it does not affect the representativeness of the SPS to be analyzed. Additionally, gene expression analysis could benefit from the use of another biological points of view. Copy

number variation could help to shed light on explaining gene expression changes for each DEG within each SPS (see widening and outliers associated with each multiclass DEG in Fig. 7). The co-integration of both omic data from the same cohort of patients could noticeably improve the downstream analysis. Finally, the relevance of the highlighted DEGs could be more widely determined by taking clinical data such as gender, race or ethnicity together with their biological involvement to certain pathways.

### ACKNOWLEDGMENT

The authors would like to thank the research group of the Institute of Bioinformatics WWU Muenster (Germany) for helpful discussions in conceptualizing this study.

### REFERENCES

- [1] A. Lomas, J. Leonardi-Bee, and F. Bath-Hextall, "A systematic review of worldwide incidence of nonmelanoma skin cancer," *Br. J. Dermatol.*, vol. 166, no. 5, pp. 1069–1080, 2012.
- [2] A. V. Giblin and J. M. Thomas, "Incidence, mortality and survival in cutaneous melanoma," *J. Plast. Reconstr. Aesthet. Surg.*, vol. 60, no. 1, pp. 32–40, 2007.
- [3] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, and D. Ioannides, "Epidemiological trends in skin cancer," *Dermatol Pract. Concept*, vol. 7, no. 2, pp. 1–6, 2017.
- [4] D. C. Whiteman, A. C. Green, and C. M. Olsen, "The growing burden of invasive melanoma: projections of incidence rates and numbers of new cases in six susceptible populations through 2031," *J. Invest. Dermatol.*, vol. 136, no. 6, pp. 1161–1171, 2016.
- [5] M. J. Eide *et al.*, "Identification of patients with nonmelanoma skin cancer using health maintenance organization claims data," *Amer. J. Epidemiol.*, vol. 171, no. 1, pp. 123–128, 2009.
- [6] C. Poupard *et al.*, "Risk of cancer in psoriasis: A systematic review and meta-analysis of epidemiological studies," *J. Eur. Acad. Dermatol.*, vol. 27, pp. 36–46, 2013.
- [7] A. Egeberg, J. P. Thyssen, G. H. Gislason, and L. Skov, "Skin cancer in patients with psoriasis," *J. Eur. Acad. Dermatol.*, vol. 30, no. 8, pp. 1349–1353, 2016.
- [8] L. Schmitz, T. Gambichler, G. Gupta, M. Stücker, and T. Dirschka, "Actinic keratosis area and severity index (AKASI) is associated with the incidence of squamous cell carcinoma," *J. Eur. Acad. Dermatol.*, vol. 32, no. 5, pp. 752–756, 2018.
- [9] B. A. Lober and C. W. Lober, "Actinic keratosis is squamous cell carcinoma," *South Med. J.*, vol. 93, no. 7, pp. 650–655, 2000.
- [10] W. C. Fix *et al.*, "MART-1-labeled melanocyte density and distribution in actinic keratosis and squamous cell cancer in situ: Pagetoid melanocytes are a potential source of misdiagnosis as melanoma in situ," *J. Cutan Pathol.*, vol. 45, no. 10, pp. 734–742, 2018.
- [11] K. B. Tan *et al.*, "Simulators of squamous cell carcinoma of the skin: diagnostic challenges on small biopsies and clinicopathological correlation," *J. Skin Cancer*, vol. 2013, 2013. [Online]. Available: <https://www.hindawi.com/journals/jsc/2013/752864/abs/>
- [12] C. K. Bichakjian *et al.*, "Merkel cell carcinoma: Critical review with guidelines for multidisciplinary management," *Cancer*, vol. 110, no. 1, pp. 1–12, 2007.
- [13] M. B. Amin *et al.*, "The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging," *CA Cancer J. Clin.*, vol. 67, no. 2, pp. 93–99, 2017.
- [14] R. S. Padilla, S. Sebastian, Z. Jiang, I. Nindl, and R. Larson, "Gene expression patterns of normal human skin, actinic keratosis, and squamous cell carcinoma: A spectrum of disease progression," *Arch. Dermatol.*, vol. 146, no. 3, pp. 288–293, 2010.
- [15] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nat Rev Genet.*, vol. 10, no. 1, pp. 57–63, 2009.
- [16] D. Bottomly *et al.*, "Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays," *PLoS One*, vol. 6, no. 3, p. e17820, 2011.
- [17] A. Sîrbu, G. Kerr, M. Crane, and H. J. Ruskin, "RNA-Seq vs dual-and single-channel microarray data: Sensitivity analysis for differential expression and clustering," *PLoS One*, vol. 7, no. 12, p. e50986, 2012.
- [18] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Res.*, vol. 33, no. 18, pp. 5914–5923, 2005.
- [19] A. Irigoyen *et al.*, "Integrative multi-platform meta-analysis of gene expression profiles in pancreatic ductal adenocarcinoma patients for identifying novel diagnostic biomarkers," *PLoS One*, vol. 13, no. 4, p. e0194844, 2018.
- [20] I. Nookaew *et al.*, "A comprehensive comparison of RNA-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*," *Nucleic Acids Res.*, vol. 40, no. 20, pp. 10084–10097, 2012.
- [21] D. Castillo *et al.*, "Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling," *BMC Bioinf.*, vol. 18, no. 1, p. 506, 2017.
- [22] D. Castillo *et al.*, "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level," *PLoS One*, vol. 14, no. 2, p. e0212127, 2019.
- [23] J. T. Leek *et al.*, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nat. Rev. Genet.*, vol. 11, no. 10, p. 733–739, 2010.
- [24] H. Gohlmann and W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays*. London, U.K.: Chapman and Hall/CRC, 2009.
- [25] Illumina and Inc., "Illumina: Illumina Gene Expression arrays." 2009.
- [26] W. W. Bin Goh, W. Wang, and L. Wong, "Why batch effects matter in omics data, and how to avoid them," *Trends Biotechnol.*, vol. 35, no. 6, pp. 498–507, 2017.
- [27] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.
- [28] C. Chen *et al.*, "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods," *PLoS One*, vol. 6, no. 2, p. e17238, 2011.
- [29] Q. Wang *et al.*, "Unifying cancer and normal RNA sequencing data from different sources," *Sci. Data*, vol. 5, p. 180061, 2018.
- [30] C. Haqq *et al.*, "The gene expression signatures of melanoma progression," *Proc. Natl. Acad. Sci. U S A*, vol. 102, no. 17, pp. 6092–6097, 2005.
- [31] J. M. Gálvez *et al.*, "Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series," *PLoS One*, vol. 13, no. 5, p. e0196836, 2018.
- [32] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS Comput. Biol.*, vol. 13, no. 5, p. e1005457, 2017.
- [33] D. Glass *et al.*, "Gene expression changes with age in skin, adipose tissue, blood and brain," *Genome Biol.*, vol. 14, no. 7, p. R75, 2013. [Online]. Available: <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-7-r75>.
- [34] B. M. Jenefer and V. Cyrilraj, "An innovative hybrid mathematical hierarchical regression model for breast cancer diseases analysis," *Cluster Comput.*, pp. 1–14, 2018.
- [35] K.-A. Lê Cao and G. J. McLachlan, "Statistical analysis on microarray data: selection of gene prognosis signatures," in *Comput. Biol.*, Springer, 2009, pp. 55–76.
- [36] J. Yang, J. Zhou, Z. Zhu, X. Ma, and Z. Ji, "Iterative ensemble feature selection for multiclass classification of imbalanced microarray data," *J. Biol. Res. (Thessalon)*, vol. 23, no. 1, 2016. [Online]. Available: <https://jbiolres.biomedcentral.com/articles/10.1186/s40709-016-0045-8>
- [37] R. A. Fisher, "The statistical utilization of multiple measurements," *Ann. Eugen.*, vol. 8, no. 4, pp. 376–386, 1938.
- [38] D. Castillo-Secilla, J. M. Gálvez, F. M. Ortuno, L. J. Herrera, and I. Rojas, "KnowSeq R/bioc package: Beyond the traditional RNA-seq pipeline. A breast cancer case study," *BMC Bioinf.*, 2019, doi: [10.21203/rs.2.16962/v1](https://doi.org/10.21203/rs.2.16962/v1)
- [39] T. Barrett *et al.*, "NCBI GEO: Mining tens of millions of expression profiles - Database and tools update," *Nucleic Acids Res.*, vol. 35, no. Suppl. 1, pp. D760–D765, 2007.
- [40] H. Parkinson *et al.*, "ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic Acids Res.*, vol. 39, no. Suppl\_1, pp. D1002–D1004, 2010.
- [41] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

- [42] W. Huber *et al.*, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nat. Methods*, vol. 12, no. 2, pp. 115–121, 2015.
- [43] D. Sean and P. S. Meltzer, "GEOquery: A bridge between the gene expression omnibus (GEO) and bioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.
- [44] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy - Analysis of affymetrix genechip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.
- [45] B. S. Carvalho and R. A. Irizarry, "A framework for oligonucleotide microarray preprocessing," *Bioinformatics*, vol. 26, no. 19, pp. 2363–2367, 2010.
- [46] P. Du, W. A. Kibbe, and S. M. Lin, "lumi: A pipeline for processing Illumina microarray," *Bioinformatics*, vol. 24, no. 13, pp. 1547–1548, 2008.
- [47] A. Kauffmann, R. Gentleman, and W. Huber, "arrayQualityMetrics - A bioconductor package for quality assessment of microarray data," *Bioinformatics*, vol. 25, no. 3, pp. 415–416, 2009.
- [48] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.
- [49] S. Anders *et al.*, "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," *Nat. Protoc.*, vol. 8, no. 9, pp. 1765–1786, 2013.
- [50] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic Acids Res.*, vol. 39, no. Suppl. 1, pp. D19–D21, 2011.
- [51] D. Kim, B. Langmead, and S. Salzberg, "HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes." 2017.
- [52] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, 2012.
- [53] H. Li *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.
- [54] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.
- [55] D. Smedley *et al.*, "The BioMart community portal: an innovative alternative to large, centralized data repositories," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W589–W598, 2015.
- [56] D. R. Zerbino *et al.*, "Ensembl 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 2017.
- [57] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.
- [58] S. Tarazona *et al.*, "Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package," *Nucleic Acids Res.*, vol. 43, no. 21, pp. e140–e140, 2015.
- [59] J. Önskog, E. Freyhult, M. Landfors, P. Rydén, and T. R. Hvidsten, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning," *BMC Bioinf.*, vol. 12, no. 1, 2011.
- [60] J. T. Leek *et al.*, "Sva: Surrogate Variable Analysis." 2018.
- [61] M. E. Ritchie *et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, 2015.
- [62] S. Mitra and Y. Hayashi, "Bioinformatics with soft computing," *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, vol. 36, no. 5, pp. 616–635, 2006.
- [63] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.
- [64] Y. Y. Ou and N. Q. K. Le, "Prediction of FAD binding sites in electron transport proteins according to efficient radial basis function networks and significant amino acid pairs," *BMC Bioinf.*, vol. 17, no. 1, 2016.
- [65] N. Q. K. Le *et al.*, "iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding," *Anal. Biochem.*, vol. 571, pp. 53–61, 2019.
- [66] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.
- [67] T. Turki and Z. Wei, "Boosting support vector machines for cancer discrimination tasks," *Comput. Biol. Med.*, vol. 101, pp. 236–249, 2018.
- [68] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 563–570, 2003.
- [69] W. S. Noble, "What is a support vector machine?," *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [70] R. M. Parry *et al.*, "k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *Pharmacogenomics J.*, vol. 10, no. 4, pp. 292–309, 2010.
- [71] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid.," in *Proc. KDD*, 1996, vol. 96, pp. 202–207.
- [72] S. Arlot *et al.*, "A survey of cross-validation procedures for model selection," *Stat Surv.*, vol. 4, pp. 40–79, 2010.
- [73] M. Stone, "Asymptotics for and against cross-validation," *Biometrika*, vol. 64, no. 1, pp. 29–35, 1977.
- [74] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Comm. Com. Inf. Sci.*, vol. 45, no. 4, pp. 427–437, 2009.
- [75] A. Fernández *et al.*, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.
- [76] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2008.
- [77] Gene Ontology Consortium, "Gene ontology consortium: Going forward," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1049–D1056, 2014.
- [78] L. Shen, L. Liu, Z. Yang, and N. Jiang, "Identification of genes and signaling pathways associated with squamous cell carcinoma by bioinformatics analysis," *Oncol Lett.*, vol. 11, no. 2, pp. 1382–1390, 2016.
- [79] E. V. Kuznetsova, E. S. Snarskaya, L. E. Zavalishina, and S. B. Tkachenko, "Immunohistochemical study of the specific features of expression of matrix metalloproteinases 1, 9 in the photoaged skin, the foci of actinic keratosis and basal cell carcinoma," *Arkh Patol.*, vol. 78, no. 6, pp. 17–22, 2016.
- [80] W. R. Swindell, M. K. Sarkar, Y. Liang, X. Xing, and J. E. Gudjonsson, "Cross-disease transcriptomics: unique IL-17A signaling in psoriasis lesions and an autoimmune PBMC signature," *J. Invest. Dermatol.*, vol. 136, no. 9, pp. 1820–1830, 2016.
- [81] Y. Sun, J. Huang, and Z. Yang, "The roles of ADAMTS in angiogenesis and cancer," *Tumour Biol.*, vol. 36, no. 6, pp. 4039–4051, 2015.
- [82] E. Kerkelä and U. Saarialho-Kere, "Matrix metalloproteinases in tumor progression: focus on basal and squamous cell skin cancer," *Exp. Dermatol.*, vol. 12, no. 2, pp. 109–125, 2003.
- [83] S. Ren *et al.*, "The impact of genomics in understanding human melanoma progression and metastasis," *Cancer Control*, vol. 15, no. 3, pp. 202–215, 2008.
- [84] C. A. Degeyses, H. B. Powell, L. B. Hsia, and B. G. Merritt, "Outcomes for invasive melanomas treated with Mohs micrographic surgery: A retrospective cohort study," *Dermatol. Surg.*, vol. 45, no. 2, pp. 223–228, 2018.
- [85] P. Gerami *et al.*, "Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma," *Clin. Cancer Res.*, vol. 21, no. 1, pp. 175–183, 2015.
- [86] J. L. Melero, S. Andrades, L. Arola, and A. Romeu, "Deciphering psoriasis. A bioinformatic approach," *J. Dermatol. Sci.*, vol. 89, no. 2, pp. 120–126, 2018.
- [87] J. Dai *et al.*, "Identification of critically carcinogenesis-related genes in basal cell carcinoma," *Oncol. Targets Ther.*, vol. 11, pp. 6957–6967, 2018.
- [88] C. M. Sauer *et al.*, "Reviewing the current evidence supporting early B-cells as the cellular origin of Merkel cell carcinoma," *Crit. Rev. Oncol. Hematol.*, vol. 116, pp. 99–105, 2017.
- [89] G. J. Kontochristopoulos, P. G. Stavropoulos, K. Krasagakis, S. Goerdt, and C. C. Zouboulis, "Differentiation between Merkel cell carcinoma and malignant melanoma: an immunohistochemical study," *Dermatology*, vol. 201, no. 2, pp. 123–126, 2000.
- [90] W. Wei *et al.*, "Identification of biomarker for cutaneous squamous cell carcinoma using microarray data analysis," *J. Cancer*, vol. 9, no. 2, pp. 400–406, 2018.
- [91] L. Q. Zheng, R. Wang, S. M. Chi, and C. X. Li, "Matrix metalloproteinase 1: A better biomarker for squamous cell carcinoma by multiple microarray analyses," *G. Ital. Dermatol. Venereol.*, vol. 154, no. 3, pp. 327–337, 2017.