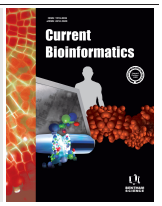
Current Bioinformatics, 2023, *18*, 40-54

RESEARCH ARTICLE

Comprehensive Pan-cancer Gene Signature Assessment through the Implementation of a Cascade Machine Learning System



Daniel Castillo-Secilla[1,2,*], Juan Manuel Galvez[2], Francisco Carrillo-Perez[2], Juan Carlos Prieto-Prieto[3], Olga Valenzuela[4], Luis Javier Herrera[2] and Ignacio Rojas[2]

[1]*Fujitsu Technology Solutions S.A., CoE Data Intelligence, Camino del Cerro de los Gamos, 1, Pozuelo de Alarcón, 28224, Madrid, Spain;* [2]*Department of Computer Architecture and Technology, University of Granada. C.I.T.I.C., Periodista Rafael Gómez Montero, 2, 18014. Granada, Spain;* [3]*Nuclear Medicine Department, IMIBIC, University Hospital Reina Sofia, Menéndez Pidal Avenue, 14004, Córdoba, Spain;* [4]*Department of Applied Mathematics, University of Granada. Facultad de Ciencias, Campus de Fuentenueva, 18071, Granada, Spain*

Abstract: ***Background*:** Despite all the medical advances introduced for personalized patient treatment and the research supported in search of genetic patterns inherent to the occurrence of its different manifestations on the human being, the unequivocal and effective treatment of cancer, unfortunately, remains as an unresolved challenge within the scientific panorama. Until a universal solution for its control is achieved, early detection mechanisms for preventative diagnosis increasingly avoid treatments, resulting in unreliable effectiveness. The discovery of unequivocal gene patterns allowing us to discern between multiple pathological states could help shed light on patients suspected of an oncological disease but with uncertainty in the histological and immunohistochemical results.

***Methods*:** This study presents an approach for pan-cancer diagnosis based on gene expression analysis that determines a reduced set of 12 genes, making it possible to distinguish between the main 14 cancer diseases.

***Results*:** Our cascade machine learning process has been robustly designed, obtaining a mean F1 score of 92% and a mean AUC of 99.37% in the test set. Our study showed heterogeneous over-orunderexpression of the analyzed genes, which can act as oncogenes or tumor suppressor genes. Upregulation of LPAR5 and PAX8 was demonstrated in thyroid cancer samples. KLF5 was highly expressed in the majority of cancer types.

***Conclusion*:** Our model constituted a useful tool for pan-cancer gene expression evaluation. In addition to providing biological clues about a hypothetical common origin of cancer, the scalability of this study promises to be very useful for future studies to reinforce, confirm, and extend the biological observations presented here. Code availability and datasets are stored in the following GitHub repository to aim for the research reproducibility: https://github.com/CasedUgr/PanCancerClassification.

Keywords: Pan-cancer, RNA-seq, TCGA, gene expression, machine learning, feature selection, CDSS.

ARTICLE HISTORY

Received: November 06, 2021
Revised: December 23, 2021
Accepted: February 09, 2022

DOI:
10.2174/1574893617666220421100512

1. INTRODUCTION

Cancer is a heterogeneous pathology that constitutes more than one hundred types of cancerous diseases, making it the second cause of death worldwide, just behind cardiovascular diseases. Although the advances in medical and diagnostic methods have allowed for improving both the early detection as well as the survival rate, the number of cancer cases increase every year. This annual increment is due to multiple factors such as pollution, diet, or even increased life expectancy. Nowadays, cancer has an immense socioeconomic impact on our society, encouraging scientists to research and discover new ways to fight against this deadly disease.

Despite its wide and heterogeneous occurrence, almost 90% of worldwide deaths and new cases of cancer in 2017 can be clustered into only 14 different cancer types, (Fig. **1**), which means that there is a minimal number of cancer types that represent an enormous incidence of cases among the global population [1]. Determining differential biological patterns from the simultaneous integrated analysis of those 14 cancer types could help massively reduce their impact in the future. With the increasing role of multi-omic analysis, the common mechanisms of emergence and development of multiple tumoral pathologies could be accurately revealed. Nowadays, a paradigm shift in treatment towards precision

*Address correspondence to this author that Fujitsu Technology Solutions S.A., CoE Data Intelligence, Camino del Cerro de los Gamos, 1, Pozuelo de Alarcón, 28224, Madrid, Spain and Department of Computer Architecture and Technology, University of Granada. C.I.T.I.C., Periodista Rafael Gómez Montero, 2, 18014. Granada, Spain; E-mail: cased@ugr.es

2212-392X/23 $65.00+.00

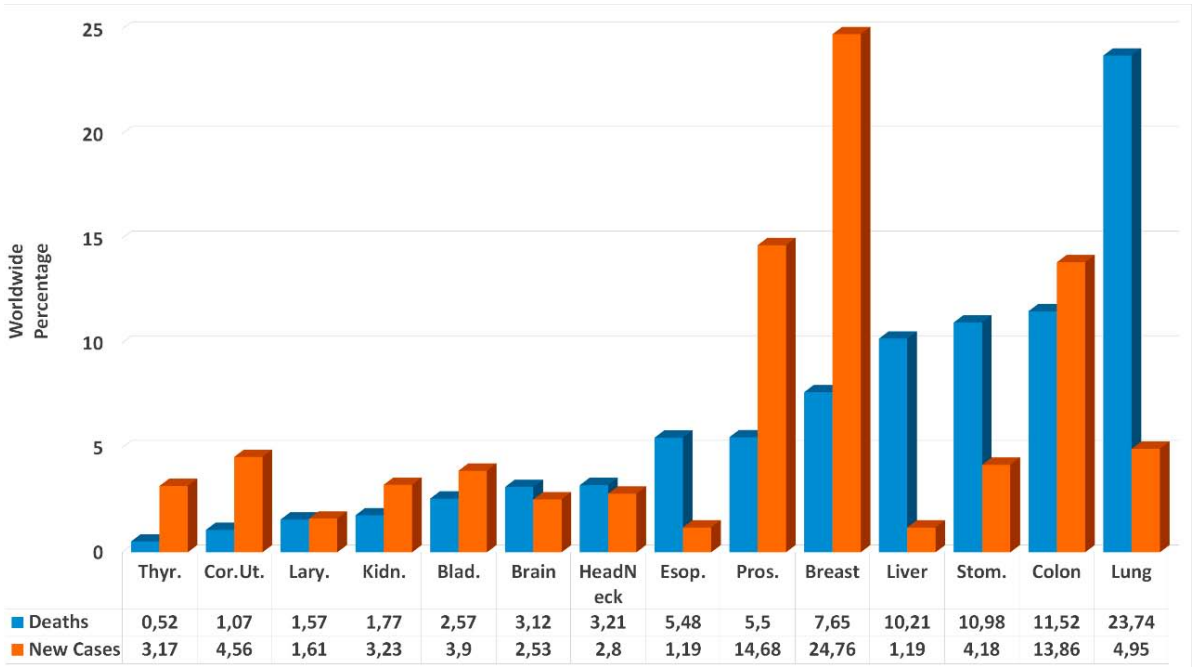


Fig. (1). Impact and incidence of the different cancer types by means of the percentage of deaths and new cases obtained in 2017. Source: Our World in Data. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

medicine has taken place. One of the main purposes of this paradigm relies on the creation and design of Clinical-Decision Support Systems (CDSS), which could help to treat each patient in a personalized way, taking into account biological information from different omics, *e.g.*, genomics, transcriptomics, proteomics or radiomics [2, 3].

The quantity of available RNA-seq data, the most renowned gene expression technology at this moment, has increased in recent years. That increase allows in-depth research studies by combining sequencing data from different cancer types. Predictive models enable the statistical validation of candidate biomarkers, especially in cases where laboratory assessment is not possible [4]. Therefore, using Machine Learning (ML) algorithms with Feature Selection (FS) techniques could improve the final detection while obtaining a subset of candidate biomarkers related to the addressed disease or diseases [5]. There are many approaches in the literature to design CDSS for specific cancer types and their subsets. Besides, interesting algorithms to deal with the heterogeneity of multiclass problems in cancer have appeared in the last few years [6, 7]. Amrane *et al*. presented a machine learning approach for classifying malignant breast cancer and benign tumors from the Breast Cancer Wisconsin Dataset, using k-Nearest Neighbour (k-NN) and Naive Bayes as classifiers and obtaining 97.51% of recognition with k-NN [8]. Alyafeai *et al*. presented a deep learning approach to detect cervical cancer using cervigrams as input [9]. The authors obtained an Area Under the Curve (AUC) of 0.82, classifying each cervix region 20 times faster than other presented models in the literature. Lu *et al*. presented a survey of the state of cancer classification using computational intelligence methods and feature selection algorithms when using gene expression data [10]. By employing unsupervised learning, Sadanandam *et al.* analyzed the gene expression of 1,290 CRC tumors, discovering associations between the clusters and the therapeutic response, the disease-free survival after surgical resection and the ability to distinguish between cell types [11]. Li *et al*. used RNA-Seq to distinguish 31 tumor types using a single data split for pan-cancer classification, correctly classifying more than 90% of test set samples using k-NN [12]. Similarly, Ma *et al*. analyzed somatic alterations in 1,699 paediatric leukemias and solid tumors across six histotypes, ultimately reporting 142 driver genes in pediatric cancers [13]. Also, using RNA-Seq data, Peng *et al*. analyzed 12 cancer types and found a 14-gene signature able to differentiate between cancerous and normal samples. By using the leave-one-out cross-validation methodology, the predictive accuracies were 92.04%, 96.23%, 91.76%, 90.05%, 88.17%, 94.29% and 99.10% for BLCA, BRCA, COAD, HNSC, LIHC, LUAD and LUSC, respectively [14]. Likewise, using miRNA-Seq data, Cheerla *et al*. presented a pan-cancer classification model and a prognosis prediction model for 21 cancer types. An accuracy of 97.2% was obtained for the pan-cancer classification task, whereas an 85% accuracy was reached for the prognosis prediction task [15]. Furthermore, previous studies in our research group have widely shown the potential of Machine Learning in different gene expression cancer problems [16-18].

Considering the previous considerations, we present a new approach to extract and validate biomarkers for a pan-cancer study formed by 6,739 samples from 14 types of different cancers and their respective control samples. This implementation receives as input data all the count files gathered from the controlled database, The Cancer Genome Atlas (TCGA) [19]. Two datasets were randomly created just at the beginning: one for training with 80% of the total samples (5,307) and another for testing with the resultant 20% (1,318)

to ensure the strictness of the process. In addition, as for the biomarker extraction and their assessment, a cross-validation strategy has been followed to avoid the emergence of dependent biomarkers only appearing within concrete partitions.

Our method takes advantage of combining a custom multiclass DEGs extraction parameter named Coverage (Cov) and the FS techniques for improving multiclass pan-cancer recognition. The multiclass gene extraction allows for retrieving DEGs that discern between different pairs of classes or cancers, giving the classifier information about all the addressed cancers due to the sum of the information provided by each DEG. In addition, FS techniques could extract a very reduced subset of Differential Expressed Genes (DEGs) with the potential to discern among the 14 different addressed cancers. This ensures a comprehensible predictive model instead of a complex model similar to a black box design. Moreover, our research includes an in-depth literature study to create a biological profile for each final DEG candidate. The whole research has been carried out by combining custom implementations with our R/Bioc framework *KnowSeq*, a flexible and modular tool encapsulating all necessary steps to perform those analyses [20]. Finally, it is important to highlight that this research includes statistical analysis to corroborate the robustness of the followed procedure and the influence of different factors on the final results. Moreover, our proposed pan-cancer tool could improve the diagnostic and therapeutic management of patients suspected of oncological disease but with indeterminate or uncertain histological and immunohistochemical results. In this sense, this tool would enable a new approach and perspective to better characterize and classify the main types of cancer worldwide from the genetic point of view.

All count files, source code and the ANOVA results are stored at the following GitHub repository: https: //github. com/CasedUgr/PanCancerClassification to support the reproducibility of this research.

2. MATERIALS AND METHODS

In this section, the implemented methodology to perform this research is explained. As depicted in Fig. (**2**), different steps can be distinguished. In order to explain the whole pipeline, the methods section is divided into three subsections: 1) Transcriptomic Pipeline; 2) DEG extraction; 3) Cascade Prediction. Furthermore, one last subsection to explain the statistical analysis designed to measure the impact of different factors on the final results is also included.

2.1. Transcriptomic Pipeline

One of the fundamental steps in transcriptomic studies relies on the correct pre-and post-processing of RAW datasets. In this sense, count files coming from the GDC portal have been previously harmonized, ensuring a significant quality for the accessible files and preventing possible BIAS from being introduced during the STAR2 alignment [21] and HTSeq read count processes [22]. Then, to calculate the gene expression given a set of count files as cancer types, Conditional Quantile Normalization (CQN) was used [23]. This normalization considers the read count of the genes and their respective gene length along with the guanine-cytosine (GC) content to carry out that normalization, returning the equivalent gene expression values for RNA-Seq samples. After this procedure, 14 gene expression sets are achieved, each set having both primary tumor and solid tissue normal samples from thecorresponding cancer.

As aforementioned, GDC count files were previously harmonized by the GDC Platform. However, this harmonization process does not consider possible remaining samples distanced from the distribution (outliers). Therefore, three different outlier detection methods were executed for each of the 14 cancer sets to reduce the possible inherent noise. It is important to highlight that the quality analysis process is based on a majority voting system, marking as outliers all those samples that have been voted at least by two of the three algorithms. The first method considers the interquartile range (IQR), one of the most famous methods to detect outliers in Machine Learning. It is based on the interquartile range, defined as the difference between the third and the first quartiles given in a distribution; any data separate from the mean more than any amount of a multiple, usually 1.5, of IQR is labeled as an outlier. For the second method, an MA plot is performed, and Hoeffding's statistic, Da, is calculated on the joint distribution of A and M for each sample. M represents the log ratio, while A represents the mean average. Finally, the Kolmogorov-Smirnov (K-S) test is also addressed, which is a nonparametric test that can be used to compare samples with a reference distribution, with the K-S statistic being the maximum vertical distance between the two distributions

In this way, our Transcriptomic Pipeline complements the GDC harmonized preprocessing, ensuring data quality for the next experimental steps performed in the study. Finally, three different datasets are created to avoid information leaks between the samples used to extract DEGs (Biclass Trn & Pancancer Trn) and samples kept for their posterior assessment (Biclass Test). The fourth dataset would be the Pan-cancer Test; however, this dataset will be formed after the first stage of the Cascade Prediction step from the well-recognized tumors.

2.2. Statistical Analysis: ANOVA

Different scenarios have been considered in developing the Analysis of Variance (ANOVA) test to better understand the configuration's influence on the classification performance (F1-Score). An ANOVA study was performed to analyze the five main factors (classifier type, feature selection method, different values of Log-fold change, COV parameter values and number of genes) (Supplementary File, Tables from **S1** to **S9** and Figures from **S1** to **S11**).

The ANOVA table in the results section shows each of the factors, the sum of squares (a measure that represents the variation or deviation with respect to the mean and calculated as a sum of the squares of the differences with respect to the mean), degrees of freedom (difference between the num-

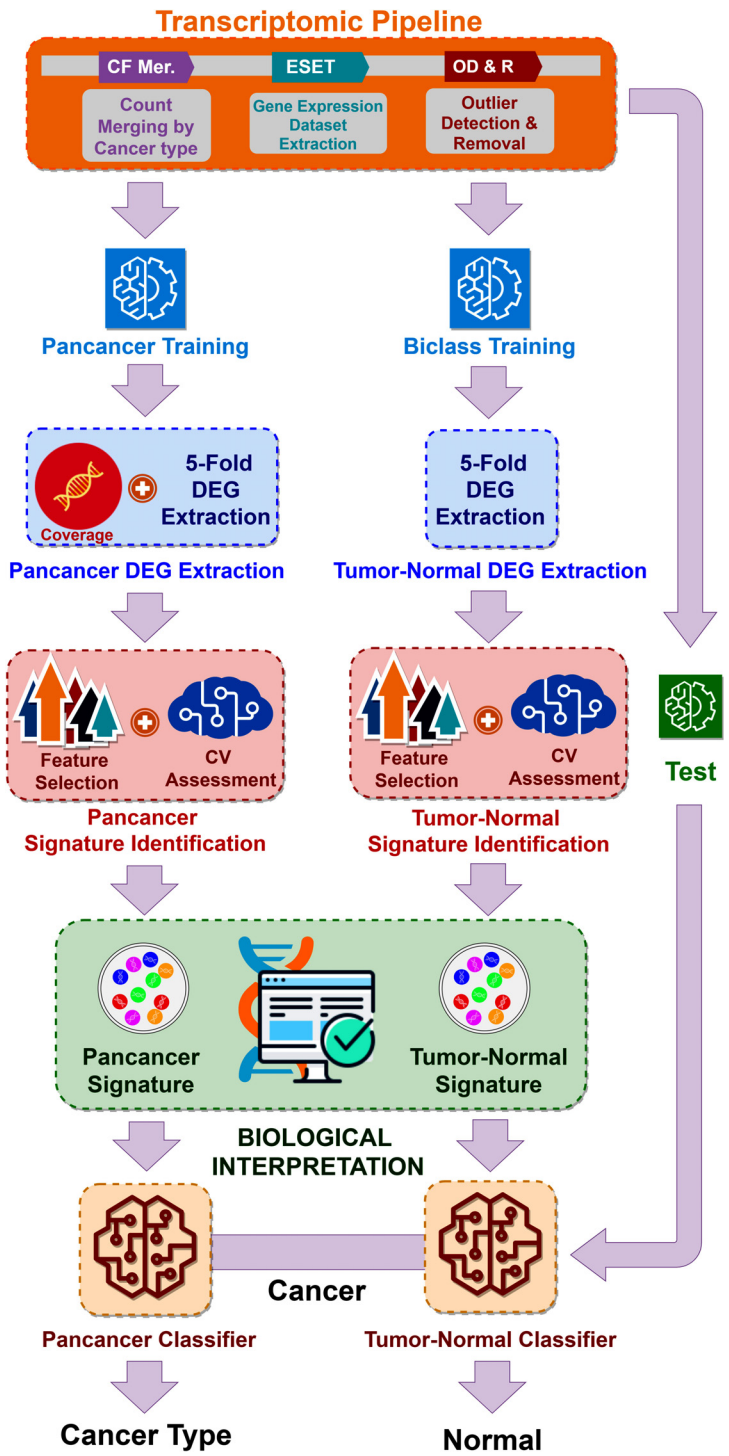


Fig (2). Flowchart of the pan-cancer pipeline implemented for our approach. Different methodological steps are specified: from raw file acquisition followed by transcriptomic analysis to Pan-cancer DEG extraction completed by whole DEG validation and assessment. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

ber of observations and parameters to be estimated), mean square (that represents an estimate of the variance of the population) and the F-Ratio (significance of the factor under study) and *p*-value (which corresponds to the lowest possible level of significance that can be chosen, for which the alternative hypothesis would still be accepted with the current observations). The F-Ratio statistic and the associated *p*-value are the most relevant ones.

2.3. DEG Extraction

Extracting potential biomarkers or features from high-resolution data is the main pillar of transcriptomics. For that reason, the epicenter of this study is the extraction of DEGs, as they could help to create a CDSS to differentiate 14 different types of cancer through Gene Expression values. During this research, two different DEG extractions were carried

out: on the one hand, a biclass biomarker extraction to obtain DEGs to differentiate between tumor or normal tissues, and on the other hand, a multiclass DEG extraction to distinguish among the 14 different types of cancer. Before starting the multiclass DEG extraction, it has to be taken intoaccount that a multiclass approach is intended to be addressed, and the classical DEG extraction methods cannot natively deal with this particularity. However, a multiclass extraction can be done by customizing the classical pipeline offered by limma [24] and DESeq2 [25]. With that consideration in mind, a multiclass parameter was proposed by Castillo *et al*. [17] named Cov to tackle multiclass problems in conjunction with limma. Cov parameter allows detecting DEGs that are differentially expressed in different classes by counting the number of biclass comparisons where this differential expression occurs, thus controlling and improving the multiclass DEG detection. It is important to note that the Cov parameter takes values between 1 (Biclass) and COV_{max}, where COV_{max} is defined in Eq. (1), with N being the number of classes in the study. Furthermore, classical threshold parameters are used, such as LFC or *P*-value.

$$Cov_{max} = \frac{N^2 - N}{2} \quad \text{where N is the number of classes} \qquad (1)$$

Although this pipeline has been designed in detail, the limitation of extracting some dependent DEGs with respect to both the biclass and pan-cancer Trn datasets can reduce the robustness. A CV DEG Selection strategy has been implemented to deal with that eventual issue. The CV technique is widely used in ML to avoid overfitting the training model step. However, this technique can be extrapolated to DEG selection by dividing both Trn Datasets into K subsets before applying the biclass and multiclass extraction. Afterward, a total of K possible gene signatures is extracted and intersected to achieve a final candidate gene signature. By applying this methodology, the final chosen DEGs do not depend only on the Trn datasets since the final DEGs are based on all evaluated partitions. Finally, two new datasets called Biclass and Pancancer Validation were created by filtering the Trn datasets with the RAW signatures calculated during this process.

2.4. Cascade Prediction

Our proposed approach has the advantage of discerning not only if a patient has cancer or not but also the concrete type of cancer among 14 possible ones. To achieve this, deployment is necessary to create two different classification stages: a first stage to identify tumors and normal tissue samples and a second stage to distinguish the concrete type of cancer for each well-classified tumor at the first stage.

2.4.1. Algorithms Overview

Since evaluating the enormous quantity of available DEG candidates at the biological level one by one can become impracticable, ML plays an important role in the statistically intelligent DEG assessment. Hence, ML helps reduce the search space to achieve a small subset of DEGs that could be evaluated in the laboratory after their intelligent assessment. Techniques such as FS or supervised classification are used to fit CDSS more than ever before. Concretely, FS has several well-known advantages when it is implemented in research of this scope:

- Models with a reduced set of features are easier to interpret and understand.
- A reduced subset of features also may dramatically shorten the training computation of the predictive models.
- Avoid the curse of dimensionality, which is a widespread problem when datasets with more features than samples are handled.
- FS is usually useful to enhance generalization and minimize overfitting since predictive models with more irrelevant or redundant features can generate models more prone to errors and value deviations.

Specifically, the mRMR algorithm has been used in this study to perform the FS, still considered one of the most powerful FS methods, originally developed for DNA Microarray by Peng *et al*. [26]. mRMR creates a ranking of features based on their relevance to the classification, penalizing the redundancy among the features. The algorithm aims to achieve the maximum relevance between a set of features X and class C but minimizes redundancy.

Our study uses supervised classification as labels are known, and we want to know, given a set of DEGs, the differences in gene expression levels existing among the different addressed states. In particular, k-NN and SVM havebeen used in this research.

K-NN is one of the most widespread and oldest supervised classifiers due to its simplicity, powerful recognition capacities, and low computational cost compared to other more complex methods [27]. Several K values are evaluated to optimize its hyperparameter, keeping the K associated with the best evaluation result as the best K to test the final models.

SVM is a complex classification and regression methodology for supervised ML published for the first time in 1995 by Cortes *et al*. [28]. It is important to highlight the relevance of the hyperparameter optimization to correctly fitthe models. In SVMs, the C hyperparameter establishes a tradeoff between the correct classification of training examples and the maximization of the decision margin.

Furthermore, the Gaussian kernel is the one used in this research. In this kernel, the σ hyperparameter manages the kernel width.

Finally, a k-fold CV strategy was implemented to improve the quality and robustness of the process. Both training and test processes are assessed in terms of accuracy, specificity, sensitivity, F1-score, receiver operating characteristic (ROC) and AUC. Accuracy represents the percentage of successes with regard to the total of samples to classify. However, this metric is unsuitable for unbalanced data or multiclass classification. Then, sensitivity or true positive

rate represents the ratio of positives that are correctly identified. Additionally, specificity or true negative rate represents the proportion of correctly identified negatives. One of the best metrics for multiclass problems is the F1 score, which also considers the classification grade for each class and not only the number of well-classified samples across all classes.

Finally, the ROC curve and AUC were also calculated to evaluate certain levels of the models. ROC curve shows the graphical relation between the sensitivity and, typically, 1-specificity. Moreover, the AUC was also calculated, considering the ROC curve. The AUC represents the model probability of classifying a random positive sample better than a random negative sample. As both metrics were designed for biclass problems, the ROC and AUC results within this manuscript were calculated by implementing a one-*vs*.-all paradigm for each class.

2.4.2. Tumor/Normal Stage

The first classification stage consists of a biclass supervised learning assessment, which uses the biclass gene signature to test the ability of those DEGs to discern between tumor and normal samples. In this sense, the available tumor and normal samples from the 14 types of cancer are grouped into the primary tumor superclass and the solid tissue normal superclass, respectively. Starting from the biclass DEGs, an FS step is performed to calculate which genes contributed with more relevant information to the classification process. As it was mentioned above, mRMR was used for this purpose. A CV step is executed after the FS to optimize the hyperparameters for each classification algorithm and select a reduced subset of DEGs with discernment capability.

Then, a biclass test classification stage is carried out with unseen samples to assess that concrete subset of DEGs. The tumor samples which have been correctly recognized will conform to the new pan-cancer dataset for the next stage of the cascade prediction methodology.

2.4.3. Pan-cancer Stage

The second classification stage uses the recognized tumors from the first stage to discern the concrete cancer type for each tumor. A random 80%-20% training/test partitioning was done. DEG extraction thresholds were imposed following the best combination of COV and LFC from the ANOVA result. Then a 5-fold DEG extraction CV was carried out using the training dataset. Finally, the common DEGs from the 5 partitions were chosen as potential pan-cancer biomarkers for assessment in the posterior ML process. An FS step was implemented to reduce the subset of DEGs from the potential biomarkers. Then, the classification algorithms were trained, and their hyperparameters were optimized. Finally, the subset of DEGs was evaluated through a test process with samples never seen before along the stage. Those final DEGs were evaluated by searching their relation in the literature with the addressed diseases or cancer in general.

3. RESULTS

3.1. Data Description

All data used in this research come from TCGA and have been acquired through the Genomic Data Commons (GDC) Portal platform. For this pan-cancer study, a total of 6,115 primary tumor samples have been gathered, belonging to 14 different types of cancers. Additionally, a total of 624 solid tissue normal samples from the different addressed cancers were also collected. Given the ratio of normal and tumor samples, it might seem that an imbalanced problem is being faced. However, primary tumor samples are divided into the 14 considered cancer types, balancing the total number of samples per class. Primary tumor samples are tumors that remain inside the affected tissue or the lymph nodes (glands), while the solid tissue normal samples are collected from the adjacent healthy tissue to the primary tumor. Table **1** contains the number of samples for each type of cancer and each tissue, together with the information about the specific number of assessed samples for training and test, shown between parentheses. Two different datasets will be taken into account to obtain robust results. The first dataset is formed by 5,328 samples (Tumor Trn + Normal Trn) and will be used to extract DEGs and train the models. The second dataset contains the 1,411 remaining samples (Tumor Test + Normal Test) and will be only used for testing the prediction capabilities of the selected DEGs. Two different datasets were formed, with an 80% proportion for training and 20% for the test.

Thanks to this division, the extracted DEGs will be independent of the samples used to assess them.

3.2. Cascade Prediction Results

Cascade prediction pipeline allows for predicting a positive case of cancer and the specific type of cancer. Our cascade implementation determines two different DEG signatures to optimize the selection of the most informative DEGs for a successful diagnosis: 1) Those DEGs indicate if the patient suffers from cancer (in general terms) or not and, if so, 2) those DEGs discerning what specific type of cancer it is. Firstly, for the biclass signature extraction, a Log_2 Fold Change (LFC) greater or equal to 1 and a p-value lower or equal to 0.001 were imposed. Then, only the first five DEGs ranked by the minimum Redundancy Maximum Relevance (mRMR) algorithm was selected for the biclass assessment as they are enough to achieve an almost total recognition of the unseen tumor samples (1,179 of 1,202 correctly classified).

The multiclass classification process presents more challenges than the biclass extraction. Moreover, multiclass extraction requires an extra parameter (COV) to retrieve truly multiclass biomarkers, making the fine-tuning process more difficult for the threshold values. An ANOVA study was conducted to assess the fine-tune parameters, as awide range of values and combinations could be considered. Under that consideration, an LFC greater or equal to 1.5, a Coverage (COV) greater or equal to 60, and a *p*-value less than or

Table 1. Collected TCGA data from the 14 different considered types of cancers for this study. The samples from both primary tumor and solid tissue normal for each cancer are shown.

Tissue/Cancer	Tumor (Trn/Test)	Normal (Trn/Test)	Total
Bladder	414 (332/82)	19 (15/4)	433
Brain	156 (125/31)	5 (4/1)	161
Breast	1059 (840/219)	122 (84/38)	1171
Colon	396 (310/86)	39 (31/8)	435
Corpus Uteri	555 (435/119)	23 (15/8)	578
Esophagus	159 (128/31)	11 (8/3)	170
Head and Neck	127 (100/27)	13 (10/3)	140
Kidney	891 (703/188)	128 (101/27)	1019
Larynx	111 (88/23)	12 (10/2)	123
Liver	408 (325/83)	58 (47/11)	466
Lung	495 (388/107)	54 (44/10)	549
Prostate	498 (398/100)	52 (42/10)	550
Stomach	344 (274/70)	30 (22/8)	374
Thyroid	502 (401/101)	58 (47/11)	560
Total	6115 (4848/1267)	624 (480/144)	6739

equal to 0.001 were chosen to extract the concrete gene signature for the experiment. Furthermore, the DEG cross-validation (CV) extraction strategy was followed to enhance the process's quality by splitting the training dataset into 5 folds, ultimately achieving five different gene signatures. Finally, the intersection of the five signatures reveals 457 DEGs independents of the 5 different folds.

After the multi-class DEG extraction process, mRMR was applied to select the most relevant genes for classification among the 457 multi-class DEGs. A CV process is implemented to profile the behavior of DEGs to design the final model and validate and tune the hyperparameters. Concretely, in our case, k-NN was chosen as there are no significant differences in the ANOVA between the two classifiers in this case and given that k-NN is simpler, more interpretable and more efficient than Support Vector Machines (SVM). For these reasons, the onlyhyperparameter to tune is the number of neighbors (K). A reduced number of DEGs was selected usingthe validation results shown in Fig. (**3**). In that figure, four different metrics (accuracy, sensitivity, specificity, and F1- Score) are shown on the X-Axis, along with the number of DEGs used to classify in the Y-axis. Moreover, the 5-fold mean and standard deviation results for each cell are drawn for each metric. Although with 3 DEGs, the accuracy reaches 82.10%, the F1-Score only reaches 15%, with a very high standard deviation. This means that there are a lot of well-classified samples; however, in multiclass terms, the recognition of the different classes is very poor. On the contrary, taking 15 DEGs ensures excellent results for the 4 metrics according to the validation model. Finally, to reduce the system complexity, the first 12 DEGs have been selected for the final model assessment as the results are very close to 15 DEG results. Having fewer DEGs than the number of classes means that each DEG allows for discerning more than two different types of cancer.

The validation step allows for profiling the model, picking up a concrete number of DEGs for the test process. After selecting the first 12 DEGs and the optimal hyperparameter value, the multiclass candidate signature was obtained. Once both the biclass and possible multiclass signatures were available, the cascade prediction process was implemented. Firstly, the test biclass dataset was assessed using the 5 biclass DEGs to differentiate between Primary Tumor and Solid Tissue Normal. Then, the correctly classified tumors were chosen as the test dataset for the multiclass stepto deduce the specific type of cancer for each tumor. Fig. (**4**) contains a cascade confusion matrix that shows the test results for both the biclass assessment with 5 DEGs and the posterior multiclass assessment with 12 DEGs. Concerning the biclass results, the upper confusion matrix concludes that 1,179 tumors of a total of 1,202 were correctly identified using our 5 proposed biclass DEGs.

Those results imply a mean accuracy equal to 96.88% and a mean F1 score equal to 98.29%, which means that a tumor iscorrectly recognized almost 100% of the time.

Then, the 1,179 tumor samples move through our hierarchy as the multiclass test dataset for the pan-cancer assessment. The multiclass results reveal a mean accuracy of 96.01% and a mean F1 score of 92.02%. This indicates that a percentage close to 100% of the tumors correctly classified at the biclass level were also correctly classified atthe multiclass level among the 14 different types of cancer. Itis important to note that larynx samples are strongly confounded with headneck samples, possibly due to the similarities between the primary tumors from these two types of cancer. In

Fig (3). Multiclass CV outcome which contains the mean results and the standard deviation for the evaluated metrics (accuracy, sensitivity, specificity, and F1-score) from 1 up to 12 DEGs. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

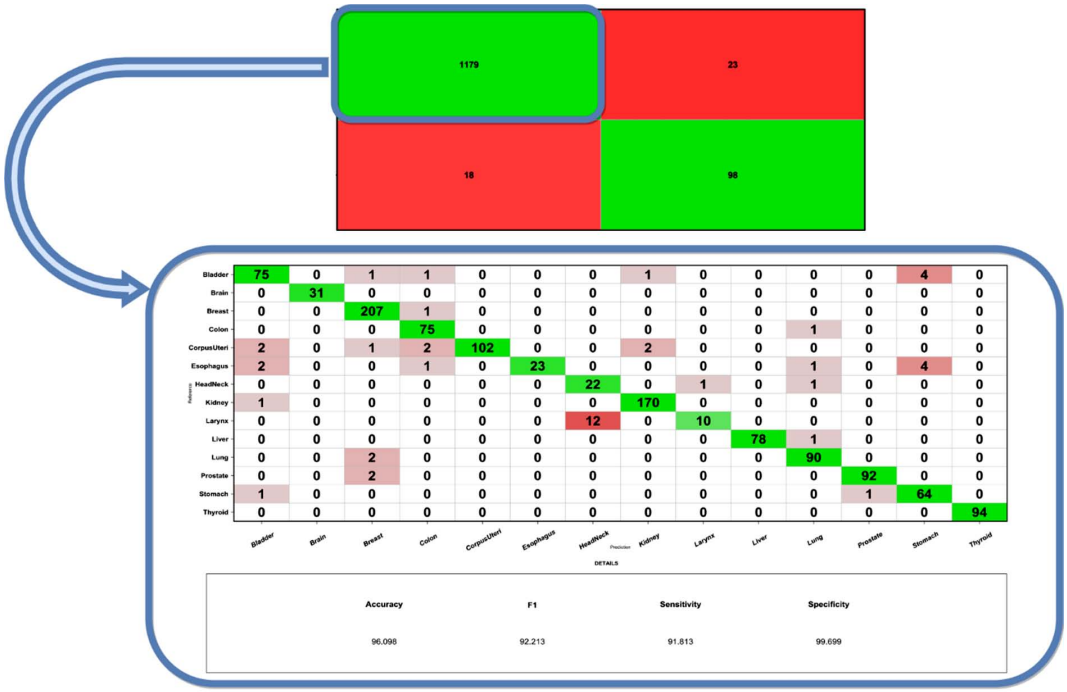


Fig (4). Combined confusion matrix which shows the biclass classification results at the background and the multi-class classification results along with the metrics at the foreground. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

fact, an extra assessment was done by combining these two tissues, and the results increased to 96.50% in the mean F1 score. Given the results, both the biclass and multiclass candidates' signatures allow for correctly predicting the state of a patient (Cancer - Normal) and specific cancer among the 14 different types of cancer addressed.

Different simulations were done to calculate the value for these metrics for each type of cancer to find both the ROC

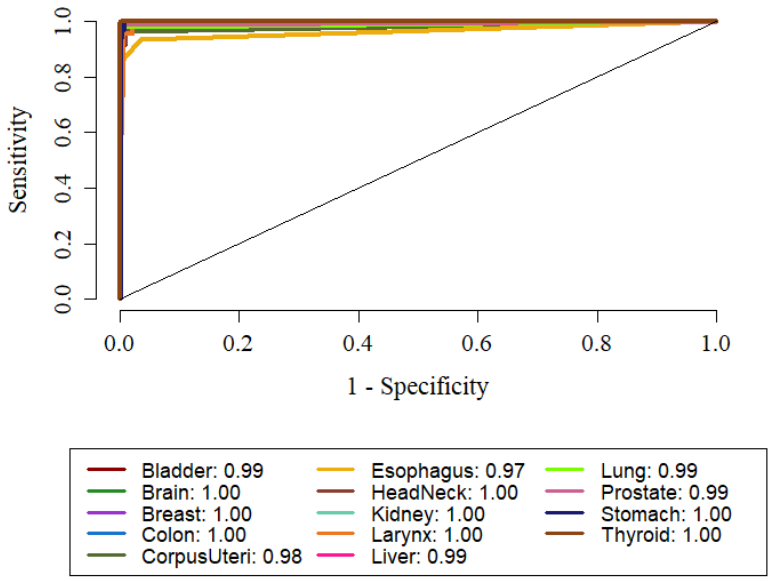


Fig (5). ROC Curve and AUC Per Classes for each of the evaluated 14 classes taking into account the classification results of each of them in comparison to the rest. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. Analysis of variance without interactions for F1-Score - Type III Sums of Squares. All F-ratios are based on the residual mean square error.

Main Effects	Sum of Squares	DF	Mean Square	F-Ratio	*P*-Value
A:Class	0,00017	1	0,00017	0,04	0,8387
B:FS	1,3956	1	1,3956	325,39	0,0000
C:LFC	0,1289	2	0,0644	15,03	0,0000
D:COV	0,1416	6	0,0236	5,50	0,0000
E:nGenes	7,3995	6	1,2332	287,53	0,0000
Residual	2,4490	571	0,0042	-	-
Total (Corrected)	11,515	587	-	-	-

curves and AUC per class for the final results. Fig. (**5**) displays the ROC curves per class and the AUC with the legend labels for each type of cancer. In this sense, all ROC curves indicate that for each type of cancer, the relationship between sensitivity and 1 - specificity ensures high performance in terms of recognition for our cascade prediction system. Furthermore, focusing on the legend of the figure, all classes reach an AUC between 98% and 100% except for the esophagus class, which reaches 96.49%. These results can be corroborated by looking at the multiclass confusion matrix in Fig. (**4**). All classes have a low number of false positives and false negatives concerning the total number of well-classified samples, except for the esophagus class. Although the larynx has more errors than the esophagus, all of them are false negatives, and there are no false positives, resulting in high AUC.

3.3. ANOVA Significance

An ANOVA has been implemented, and the impact of different factors on the final F1 score is achieved. There are five factors to assess: Classification Algorithm, Feature Selector, LFC, Cov and Number of Genes. Several simulations were done to ensure robust results by combining many values for the different factors. Although the most important considerations are shown herein, the whole ANOVA study has been included in the Pan-cancer GitHub Repository indicated at the end of the Introduction Section.

F1-score results from ANOVA without interactions reveal strong statistical differences when different configurations of the evaluated factors are implemented. Concretely, as Table **2** shows, the Classification Algorithm (k-NN or SVM) is the only factor that does not have a strong influence on the F1-score. Conversely, the other four assessed factors influence the F1-score achieved by the final models in a significant manner as they have a *p*-value lower than 0.05.

4. DISCUSSION

A cascade strategy was carefully designed to predict when a patient has cancer and specific cancer among 14 different types. Classification outcomes at the tumor *vs*. the

normal stage could discern almost all tumor samples, employing a reduced set of 5 DEGs. Considering the low number of DEGs needed, these 5 DEGs seem to have a strong discernment capability between tumor and normal. Furthermore, Fig. (**6**) shows the gene expression level for each of the 5 DEGs between the addressed states, revealing notable differences. Focusing on the ANOVA test without interaction reveals statistical significance between most of the evaluated factors and the achieved F1 score. The only non-significant factor was the ML algorithm employed to classify the samples. The ANOVA test verifies that, in this case, both evaluated algorithms (k-NN and SVM) do not have statistical differences between them, supporting that interpretation as an indicator of the robustness of our proposed DEG extraction methodology. Finally, an in-depth study of the literature for the 5 DEGs was carried out to profile each of them properly.

MELK (Maternal Embryonic Leucine Zipper Kinase) encodes a serine/threonine protein kinase that regulates the mitotic cycle, apoptosis, and splicing. Several studies correlate MELK overexpression in multiple cancer types [29]. It has been associated with BCL2L14 phosphorylation, thus playing a significant role in breast cancer [30], especially in the triple-negative variant [31]. The implication in endometria [32] and natural killer/T-cell lymphoma [33] has also been reported.

Dermatopontin (DPT) codifies an extracellular matrix (ECM) protein expressed throughout the organism, regulating cellular interaction and attachment. Furthermore, it has been postulated that it activates Transforming Growth Factor Beta (TFG-β), thus inhibiting cell proliferation. Hepatocellular carcinoma [34], oral cancer [35], as well as papillary thyroid cancer [36] are examples of tumors whose proliferation is regulated by DPT.

The SOSTDC1 gene is a member of sclerostin. The protein it encodes acts as a bone morphogenetic protein antagonist, regulating cellular proliferation, differentiation, and apoptosis. SOSTDC1 expression has recently been demonstrated to inhibit bone metastasis in non-small cell lung cancer [37], follicular thyroid cancer [38], and gastric cancer progression [39]. On the other hand, its downregulation reduces metastatic liver invasion in colorectal cancer [40].

Homo sapiens dishevelled binding antagonist of beta-catenin 2 (DACT2), as a member of the catenin family, is involved in cell adhesion and regulation of intracellular signaling pathways. Several cancer types have been associated with DACT2 downregulation, including lung adenocarcinoma [41], colorectal cancer [42], and breast cancer [43, 44], all of them through Wnt-β-catenin signaling pathway regulation.

LYVE1 or lymphatic vessel endothelial hyaluronan receptor 1 codifies a type I integral membrane glycoprotein, which may mediate lymphatic hyaluronan transport and tumor metastasis. LYVE1 overexpression has been demonstrated in lung cancer [45] and breast cancer [46, 47] by angiogenesis and lymphangiogenesis mechanisms.

After the first stage, the multiclass stage begins taking into account the previously classified tumors, which are now used to discern the specific type for each. Concretely, a subset of 12 DEGs (less than 1 DEG per assessed class) was extracted and evaluated with the ML step. Fig. (**4**) reveals the precision of our methodology, reaching a total of 92% F1- Score in the 14-class problem. Concretely, the only 2 cancers confounded were the larynx and head-neck.

Although the results in the manuscript split these 2 classes, the F1 score increased to 96% when they were treated as one. It could be interesting to evaluate in-depth similarities in the gene expression level of these 2 types of cancer to bring light to this problem. Focusing on Fig. (**7**), which shows the multiclass boxplots for each DEG, it can be seen how each DEG allows for discerning between two or more cancers.

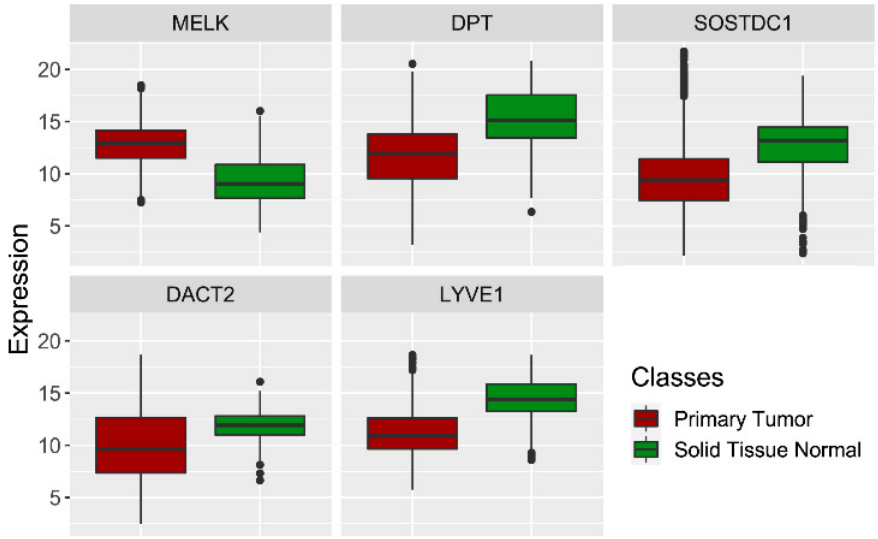


Fig (6). Boxplots of the final 5 DEGs selected at the biclass stage. Each boxplot shows the gene expression level differences between the tumor and normal samples. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

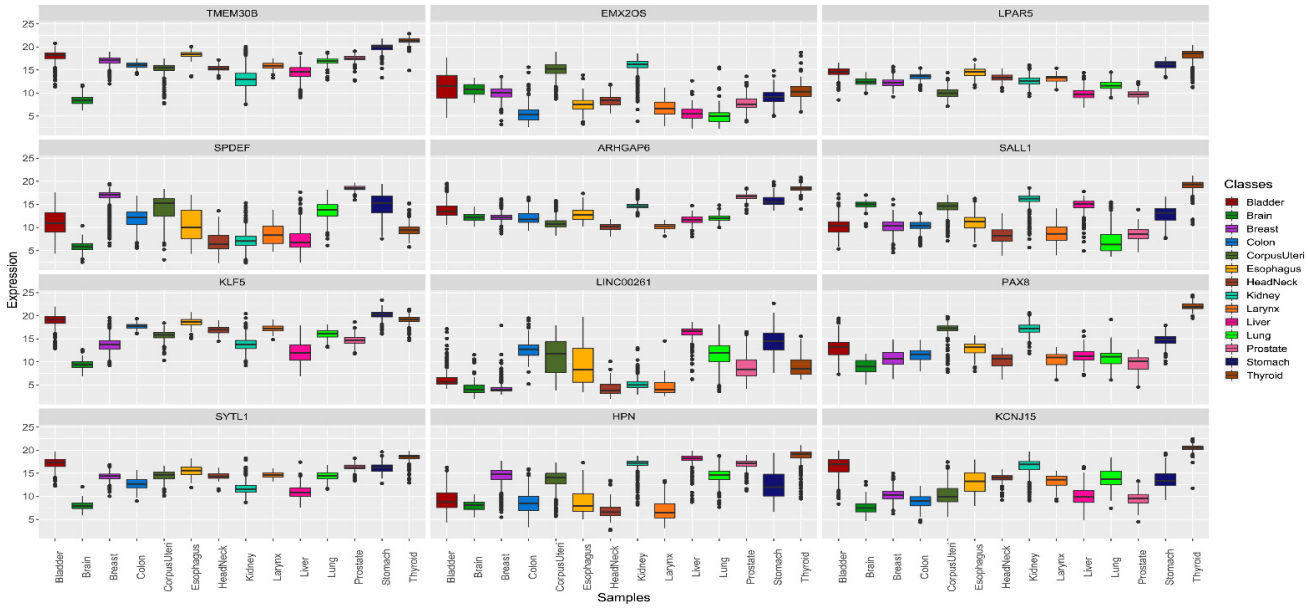


Fig (7). Boxplots of the final 5 DEGs selected at the biclass stage. Each boxplot shows the differences in gene expression level between both the tumor and normal samples. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

That extraction was thanks to our COV parameter, which extracts robust multiclass DEGs. Moreover, as for the biclass stage DEGs, an exhaustive study was done for each gene to determine the relation of this subset with the different addressed cancers.

In order to visualize if the found DEGs were grouping together samples from the same class, the T-Distributed Stochastic Neighbor embedding (T-SNE) was used. T-SNE is a commonly used visualization technique that maps points in high-dimensional vector spaces into lower dimensions. It converts vector similarities into joint probabilities, differentiating it from other dimensionality reduction techniques [48]. By doing so, it allows for generating visually distinct clusters that represent patterns in the data. The representation for the 14 specific cancer types using the final 12 DEGs chosen and all the available samples (training + test) can be observed in Fig. (**8**). The clusters reveal how almost all cancer types were clearly separated, with strong overlap between the larynx and the headneck samples.

TMEM30B encodes a transmembrane protein expressed as a component of cellular and mitochondrial membranes. It belongs to the CDC50/LEM3 family, involved in protein binding and aminophospholipid transport processes. Its downregulation of methylation mechanisms has been associated with poor prognosis in malignant mesothelioma [49]. Underexpression may have a similar impact on clear cellrenal carcinoma [50] and be related to the progression and/or recurrence of meningiomas [51].

EMX2 opposite strand/antisense RNA (EMX2OS) has been reported as a potential oncogene for human cancer. It is highly overexpressed in ovarian cancer cells [52], and its downregulation has recently been associated with worse outcomes in clear cell renal carcinoma [53].

LPAR5 codifies a G protein-coupled transmembrane receptor similar to rhodopsin (a lysophosphatidic acid receptor). It is overexpressed in thyroid tumor cells, constituting a promising therapeutic target related to the mTOR pathway [54].

Prostate-derived ETS factor SPDEF has been proposed to function as an androgen-independent transactivator of the prostate-specific antigen (PSA) promoter. Meiners *et al*. [55] proved in a retrospective study including over 12,000 microarray samples that SPDEF upregulation indicated poor prognosis in prostate cancer (higher Gleason score andrisk of biochemical PSA recurrence). On the other hand, Ye *et al*. [56] suggested that it may play a dual role in breast cancer, helping tumor development but inhibiting tumor progression by modulating its microenvironment at the same time.

ARHGAP6 (Rho GTPase-activating protein 6) acts as a regulator of actin polymerisation, involving multiple cellular processes related to cytoskeletal actin remodeling. CLDN18-ARHGAP fusion has been reported to significantly decrease survival rates in gastric cancer [57]. Moreover, ARHGAP6 expression inhibited proliferation and invasion in cervical [58], breast [59], as well as lung cancer cells [60], becoming a potential therapeutic target as a tumor suppressor gene.

Spalt-like transcription factor 1 (SALL1) is a member of the Krüppel-associated box (KRAB) family, which contains zinc finger proteins and can modulate organogenesis. Chi *et al*. [61] demonstrated that its overexpression reduces glioma cell proliferation and migration through the Wnt/ß-catenin pathway. Furthermore, Ma *et al*. [62] proved that it also plays a significant role as a senescence inducer in breast tumor tissues, being downregulated in these malignant cells. Li *et al.* [63] concluded that miR-4286 knockdown inhibits prostate cancer proliferation through SALL1 signaling.

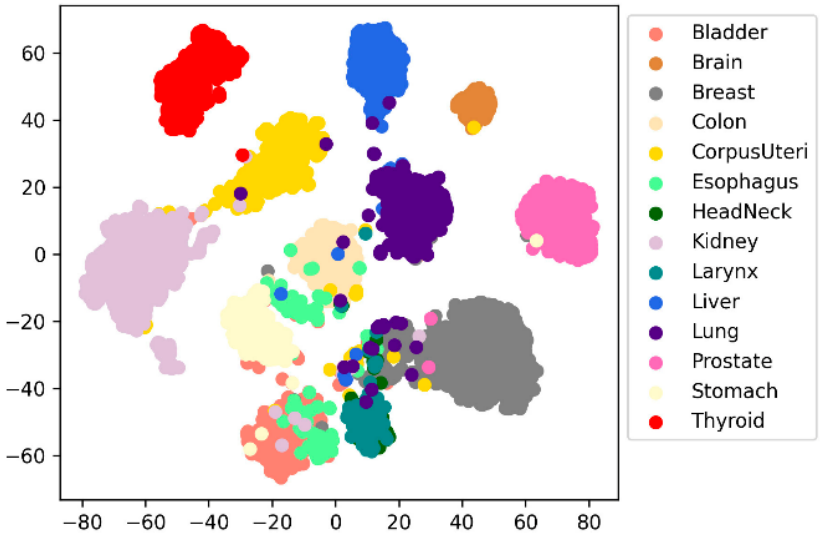


Fig (8). T-SNE-mapped representations of the final pancancer gene signature (12 genes) for all samples in the dataset. Well defined clusters can be observed for each class, with some overlapping in those that are more difficult to distinguish by the classifier. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

KLF5 or Krüppel-like factor 5 is another zinc-finger transcription factor family member, implicated in multiple signaling pathways such as Wnt, Ras, Hippo, or Notch, and isan excellent candidate as a therapeutic oncologic target [64]. Recently Chen *et al*. [65] determined that KLF5 was highly overexpressed in gastric cancer cells by regulating p21 and CDK4. Additionally, ubiquitin-specific protease 3 (USP3) has been demonstrated to induce upregulation of KLF5, thus promoting tumorigenesis in breast cancer [66].

Homo sapiens long intergenic non-protein coding RNA 261 (LINC00261) is a long noncoding RNA that acts as a negativecell growth regulator. Its overexpression suppresses lung cancer metastasis proliferation *via* FOXO1 [67] as well as colorectal cancer progression through the Wnt/β-catenin pathway [68] and pancreatic cancer invasion *via* c-Myc methylation [69].

PAX8 codifies a paired box family member involved in thyroid follicular cell growth, being widely studied because of its direct implications in thyroid carcinoma due to PAX8/PPARY chromosomal rearrangement [70]. Noncoding somatic mutations may represent a relevant phenomenon in ovarian cancer physiopathology [71], and it has been determined that PAX8 overexpression was associated with poorer prognosis in gastric cancer in combination with SOX18 [72].

Synaptotagmin-like protein 1 (SYTL1) encodes a vesicle trafficking protein which may be closely related to cytotoxic granule exocytosis in lymphocytes mediated by Ras-related protein Rab-27A RAB27A. SYTL1 activation by MEIS1 induces leukemogenesis and interaction between leukemic cells/bone marrow [73]. Ho *et al*. [74] stated that SYTL1 was significantly downregulated in FGFR3-non-mutated bladder tumor samples at higher tumor invasion grades (muscular- invasion stage).

HPN codifies the serine protease hepsin, a type II transmembrane serine protease related to coagulation mechanism and cellular morphology. Overexpression of HPN predicts poorer outcomes in gastric cancer [75], but most importantly, several studies demonstrated the correlation between alterations in HPN gen and a higher risk and viability of prostate cancer [76, 77], even though this serine protease is also histochemically expressed in normal prostatictissues [78].

Finally, KCNJ15 encodes ATP-sensitive inward-rectifier potassium channel 15, a membrane protein that facilitates the incorporation of potassium ions into the cytoplasm. Tissue overexpression has recently been associated with poor prognosis in esophageal squamous cell carcinoma [79] and renal carcinoma [80].

CONCLUSION

Along this research, a 14-cancer type differentiation at the gene expression level has been addressed by implementing a cascade machine learning strategy. The designed pipeline's main goal is to make understanding the underlying differential expressed gene information more comprehensive. With this goal in mind, the proposed cascade methodology allows for extracting and assessing a reduced set of differential expressed genes with the potential to discern not only those patients who have cancer butalso the specific type of cancer among the 14 considered (which represent 90% of the total diagnosed cancer worldwide). Concretely, both the biclass and multiclass stagesachieve an F1 score of 98% and 92%, respectively, achieving almost total recognition for the handled cancer types. Furthermore, the evaluation of the AUCROC metric for eachtype of cancer supports the robustness of our methodology.

Moreover, the posterior analysis of the biclass and multiclass identified genes showed a certain level of heterogeneity in terms of over- and underexpression, acting either

as oncogenes or tumor suppressor genes. This was supported by the literature revision and verified in the boxplot gene expression levels and the T-SNE clusters. In addition, these genes are involved in multiple oncogenic metabolic pathways, resulting in a high expression pattern in almost all tumor classes. Considering the previous considerations, the cross-validation of the differential expressed genes extraction, along with the coverage multiclass parameter, allows for retrieving genes involved in several cancerous processes.

In summary, our machine learning model represents a solid approach for a comprehensible pan-cancer cross-evaluation system implying robust gene expression extraction through RNA-Seq and its significance in oncology. However, further prospective studies are required to better understand tumour proliferation, invasion, and dissemination risk at a genetic level.

LIST OF ABBREVIATIONS

CDSS	=	Clinical-Decision Support Systems
ML	=	Machine Learning
FS	=	Feature Selection
k-NN	=	k-Nearest Neighbour
AUC	=	Area Under the Curve

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

A public repository was created to facilitate the replication of the pipeline. R-codes were used to carry out the experiments, and the preprocessed dataset was uploaded to the GitHub repository. It can be found at https://github.com/CasedUgr/PanCancerClassification.

FUNDING

The Spanish Ministry of Sciences, Innovation and Universities funded this work under the projects RTI2018-101674-B-I00 and PID2021-128317OB-I00 in collaboration with the Government of Andalusia under the projects P20_00163 and CV20-64934.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Our world in data. 2020. Cancer deaths by type. Available from: https://ourworldindata.org/grapher/cancer-deaths-

[2] Gómez-López G, Dopazo J, Cigudosa JC, Valencia A, Al-Shahrour F. Precision medicine needs pioneering clinical bioinformaticians. Brief Bioinform 2019; 20(3): 752-66. http://dx.doi.org/10.1093/bib/bbx144 PMID: 29077790

[3] Beauchemin M, Murray MT, Sung L, Hershman DL, Weng C, Schnall R. Clinical decision support for therapeutic decision-making in cancer: A systematic review. Int J Med Inform 2019; 130: 103940. http://dx.doi.org/10.1016/j.ijmedinf.2019.07.019 PMID: 31450082

[4] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014; 13: 8-17. http://dx.doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696

[5] Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med 2019; 112: 103375. http://dx.doi.org/10.1016/j.compbiomed.2019.103375 PMID: 31382212

[6] Ren X, Wang Y, Chen L, Zhang XS, Jin Q. ellipsoidFN: A tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. Nucleic Acids Res 2013; 41(4): e53-3. http://dx.doi.org/10.1093/nar/gks1288 PMID: 23262226

[7] Zou M, Duren Z, Yuan Q, *et al.* MIMIC: An optimization method to identify cell type-specific marker panel for cell sorting. Brief Bioinformatics 2021; 22. (6): bbab235. http://dx.doi.org/10.1093/bib/bbab235

[8] Amrane M, Oukid S, Gagaoua I. Breast cancer classification using machine learning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). In: IEEE; 2018; pp. 1-4.

[9] Alyafeai Z, Ghouti L. A fully-automated deep learning pipeline for cervical cancer classification. Expert Syst Appl 2020; 141: 112951. http://dx.doi.org/10.1016/j.eswa.2019.112951

[10] Lu Y, Han J. Cancer classification using gene expression data. Inf Syst 2003; 28(4): 243-68. http://dx.doi.org/10.1016/S0306-4379(02)00072-8

[11] Sadanandam A, Lyssiotis CA, Homicsko K, *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med 2013; 19(5): 619-25. http://dx.doi.org/10.1038/nm.3175 PMID: 23584089

[12] Li Y, Kang K, Krahn JM, *et al.* A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017; 18(1): 508. http://dx.doi.org/10.1186/s12864-017-3906-0 PMID: 28673244

[13] Ma X, Liu Y, Liu Y, *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature 2018; 555(7696): 371-6. http://dx.doi.org/10.1038/nature25795 PMID: 29489755

[14] Peng L, Bian XW, Li DK, *et al.* Large-scale rna-seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 tcga cancer types. Sci Rep 2015; 5(1): 13413. http://dx.doi.org/10.1038/srep13413 PMID: 26292924

[15] Cheerla N, Gevaert O. Microrna based pan-cancer diagnosis and treatment recommendation. BMC Bioinformatics 2017; 18(1): 32. http://dx.doi.org/10.1186/s12859-016-1421-y PMID: 28086747

[16] Castillo D, Gálvez JM, Herrera LJ, Román BS, Rojas F, Rojas I. Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. BMC Bioinformatics 2017; 18(1): 506. http://dx.doi.org/10.1186/s12859-017-1925-0 PMID: 29157215

[17] Castillo D, Galvez JM, Herrera LJ, *et al.* Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. PLoS One 2019; 14(2): e0212127. http://dx.doi.org/10.1371/journal.pone.0212127 PMID: 30753220

[18] Gálvez JM, Castillo-Secilla D, Herrera LJ, *et al.* Towards improving skin cancer diagnosis by integrating microarray and rna- seq datasets. IEEE J Biomed Health Inform 2020; 24(7): 2119-30. http://dx.doi.org/10.1109/JBHI.2019.2953978 PMID: 31871000

[19] Weinstein JN, Collisson EA, Mills GB, *et al.* Cancer Genome Atlas Research Network. The cancer genome atlas pan- cancer analysis project. Nat Genet 2013; 45(10): 1113-20. http://dx.doi.org/10.1038/ng.2764 PMID: 24071849

[20] Castillo-Secilla D, Gálvez JM, Carrillo-Perez F, *et al.* KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. Comput Biol Med 2021; 133: 104387. http://dx.doi.org/10.1016/j.compbiomed.2021.104387 PMID: 33872966

[21] Dobin A, Davis CA, Schlesinger F, *et al.* STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29(1): 15-21. http://dx.doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

[22] Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015; 31(2): 166-9. http://dx.doi.org/10.1093/bioinformatics/btu638 PMID: 25260700

[23] Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 2012; 13(2): 204-16. http://dx.doi.org/10.1093/biostatistics/kxr054 PMID: 22285995

[24] Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43(7): e47-7. http://dx.doi.org/10.1093/nar/gkv007 PMID: 25605792

[25] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15(12): 550. http://dx.doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

[26] Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005; 27(8): 1226-38. http://dx.doi.org/10.1109/TPAMI.2005.159 PMID: 16119262

[27] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967; 13(1): 21-7. http://dx.doi.org/10.1109/TIT.1967.1053964

[28] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20(3): 273-97. http://dx.doi.org/10.1007/BF00994018

[29] Gray D, Jubb AM, Hogue D, *et al.* Maternal embryonic leucine zipper kinase/murine protein serine-threonine kinase 38 is a promising therapeutic target for multiple cancers. Cancer Res 2005; 65(21): 9751-61. http://dx.doi.org/10.1158/0008-5472.CAN-04-4531 PMID: 16266996

[30] Lin M-L, Park J-H, Nishidate T, Nakamura Y, Katagiri T. Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family. Breast Cancer Res 2007; 9(1): R17. http://dx.doi.org/10.1186/bcr1650 PMID: 17280616

[31] Pitner MK, Taliaferro JM, Dalby KN, Bartholomeusz C. MELK: A potential novel therapeutic target for TNBC and other aggressive malignancies. Expert Opin Ther Targets 2017; 21(9): 849-59. http://dx.doi.org/10.1080/14728222.2017.1363183 PMID: 28764577

[32] Xu Q, Ge Q, Zhou Y, *et al.* MELK promotes Endometrial carcinoma progression *via* activating mTOR signaling pathway. EBioMedicine 2020; 51: 102609. http://dx.doi.org/10.1016/j.ebiom.2019.102609 PMID: 31915116

[33] Li B, Yan J, Phyu T, *et al.* MELK mediates the stability of EZH2 through site-specific phosphorylation in extranodal natural killer/T-cell lymphoma. Blood 2019; 134(23): 2046-58. http://dx.doi.org/10.1182/blood.2019000381 PMID: 31434700

[34] Liu S, Qiu J, He G, *et al.* Dermatopontin inhibits WNT signaling pathway *via* CXXC finger protein 4 in hepatocellular carcinoma. J Cancer 2020; 11(21): 6288-98. http://dx.doi.org/10.7150/jca.47157 PMID: 33033513

[35] Yamatoji M, Kasamatsu A, Kouzu Y, *et al.* Dermatopontin: A potential predictor for metastasis of human oral cancer. Int J Cancer 2012; 130(12): 2903-11. http://dx.doi.org/10.1002/ijc.26328 PMID: 21796630

[36] Guo Y, Li H, Guan H, *et al.* Dermatopontin inhibits papillary thyroid cancer cell proliferation through MYC repression. Mol Cell Endocrinol 2019; 480: 122-32. http://dx.doi.org/10.1016/j.mce.2018.10.021 PMID: 30391671

[37] Chen G, Gong H, Wang T, *et al.* SOSTDC1 inhibits bone metastasis in non-small cell lung cancer and may serve as a clinical therapeutic target. Int J Mol Med 2018; 42(6): 3424-36. http://dx.doi.org/10.3892/ijmm.2018.3926 PMID: 30320379

[38] Zhou Q, Chen J, Feng J, Xu Y, Zheng W, Wang J. SOSTDC1 inhibits follicular thyroid cancer cell proliferation, migration, and EMT *via* suppressing PI3K/Akt and MAPK/Erk signaling pathways. Mol Cell Biochem 2017; 435(1-2): 87-95. http://dx.doi.org/10.1007/s11010-017-3059-0 PMID: 28551845

[39] Cui Y, Zhang F, Jia Y, *et al.* The BMP antagonist, SOSTDC1, restrains gastric cancer progression *via* inactivation of c-Jun signaling. Am J Cancer Res 2019; 9(11): 2331-48. PMID: 31815038

[40] Bartolomé RA, Pintado-Berninches L, Jaén M, de Los Ríos V, Imbaud JI, Casal JI. SOSTDC1 promotes invasion and liver metastasis in colorectal cancer *via* interaction with ALCAM/CD166. Oncogene 2020; 39(38): 6085-98. http://dx.doi.org/10.1038/s41388-020-01419-4 PMID: 32801337

[41] Zhang N, Li Y, Xie M, *et al.* DACT2 modulated by TFAP2A-mediated allelic transcription promotes EGFR-TKIs efficiency in advanced lung adenocarcinoma. Biochem Pharmacol 2020; 172: 113772. http://dx.doi.org/10.1016/j.bcp.2019.113772 PMID: 31866302

[42] Lu L, Wang Y, Ou R, *et al.* Dact2 epigenetic stimulator exerts dual efficacy for colorectal cancer prevention and treatment. Pharmacol Res 2018; 129: 318-28. http://dx.doi.org/10.1016/j.phrs.2017.11.032 PMID: 29199082

[43] Li J, Zhang M, He T, *et al.* Methylation of DACT2 promotes breast cancer development by activating Wnt signaling. Sci Rep 2017; 7(1): 3325. http://dx.doi.org/10.1038/s41598-017-03647-3 PMID: 28607412

[44] Guo L, Wang X, Yang Y, *et al.* Methylation of DACT2 contributes to the progression of breast cancer through activating WNT signaling pathway. Oncol Lett 2018; 15(3): 3287-94. PMID: 29435071

[45] Li P, Cong Z, Qiang Y, *et al.* Clinical significance of CCBE1 expression in lung cancer. Mol Med Rep 2018; 17(2): 2107-12. PMID: 29207117

[46] Van der Auwera I, Van den Eynden GG, Colpaert CG, *et al.* Tumor lymphangiogenesis in inflammatory breast carcinoma: A histomorphometric study. Clin Cancer Res 2005; 11(21): 7637-42. http://dx.doi.org/10.1158/1078-0432.CCR-05-1142 PMID: 16278382

[47] Hunter S, Nault B, Ugwuagbo KC, Maiti S, Majumder M. Mir526b and mir655 promote tumour associated angiogenesis and lymphangiogenesis in breast cancer. Cancers (Basel) 2019; 11(7): 938. http://dx.doi.org/10.3390/cancers11070938 PMID: 31277414

[48] Van der Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res 2008; 9(11): 2579-605.

[49] Hama R, Watanabe Y, Shinada K, *et al.* Characterization of DNA hypermethylation in two cases of peritoneal mesothelioma. Tumour Biol 2012; 33(6): 2031-40. http://dx.doi.org/10.1007/s13277-012-0462-8 PMID: 22836805

[50] Wrzesiński T, Szelag M, Cieślikowski WA, *et al.* Expression of pre-selected TMEMs with predicted ER localization as potential classifiers of ccRCC tumors. BMC Cancer 2015; 15(1): 518. http://dx.doi.org/10.1186/s12885-015-1530-4 PMID: 26169495

[51] Pérez-Magán E, Campos-Martín Y, Mur P, *et al.* Genetic alterations associated with progression and recurrence in meningiomas. J Neuropathol Exp Neurol 2012; 71(10): 882-93. http://dx.doi.org/10.1097/NEN.0b013e31826bf704 PMID: 22964784

[52] Duan M, Fang M, Wang C, Wang H, Li M. Lncrna emx2os induces proliferation, invasion and sphere formation of ovarian cancer cells *via* regulating the mir-654-3p/akt3/pd-l1 axis. Cancer Manag Res 2020; 12: 2141-54. http://dx.doi.org/10.2147/CMAR.S229013 PMID: 32273754

[53] Jiang H, Chen H, Wan P, Song S, Chen N. Downregulation of enhancer RNA EMX2OS is associated with poor prognosis in kidney renal clear cell carcinoma. Aging (Albany NY) 2020; 12(24): 25865-77. http://dx.doi.org/10.18632/aging.202151 PMID: 33234727

[54] Wu C-Y, Zheng C, Xia E-J, *et al.* Lysophosphatidic acid receptor 5 (lpar5) plays a significance role in papillary thyroid cancer *via* phosphatidylinositol 3-kinase/akt/mammalian target of rapamycin (mtor) pathway. Med Sci Mon 2020; 26: e919820. http://dx.doi.org/10.12659/MSM.919820

[55] Meiners J, Schulz K, Möller K, *et al.* Upregulation of SPDEF is associated with poor prognosis in prostate cancer. Oncol Lett 2019; 18(5): 5107-18. http://dx.doi.org/10.3892/ol.2019.10885 PMID: 31612022

[56] Ye T, Feng J, Wan X, Xie D, Liu J. Double agent: Spdef gene with both oncogenic and tumor-suppressor functions in breast cancer. Cancer Manag Res 2020; 12: 3891-902. http://dx.doi.org/10.2147/CMAR.S243748 PMID: 32547225

[57] Zhang W-H, Zhang S-Y, Hou Q-Q, *et al.* The significance of the cldn18-arhgap fusion gene in gastric cancer: A systematic review and meta-analysis. Front Oncol 2020; 10: 1214. http://dx.doi.org/10.3389/fonc.2020.01214 PMID: 32983960

[58] Li J, Liu Y, Yin Y. Inhibitory effects of Arhgap6 on cervical carcinoma cells. Tumour Biol 2016; 37(2): 1411-25. http://dx.doi.org/10.1007/s13277-015-4502-z PMID: 26628301

[59] Chen W-X, Lou M, Cheng L, *et al.* Bioinformatics analysis of potential therapeutic targets among *ARHGAP* genes in breast cancer. Oncol Lett 2019; 18(6): 6017-25. http://dx.doi.org/10.3892/ol.2019.10949 PMID: 31788076

[60] Wu Y, Xu M, He R, Xu K, Ma Y. ARHGAP6 regulates the proliferation, migration and invasion of lung cancer cells. Oncol Rep 2019; 41(4): 2281-888. http://dx.doi.org/10.3892/or.2019.7031 PMID: 30816546

[61] Chi D, Zhang W, Jia Y, Cong D, Hu S. Spalt-like transcription factor 1 (sall1) gene expression inhibits cell proliferation and cell migration of human glioma cells through the wnt/β- catenin signaling pathway. Med Sci Monit Basic Res 2019; 25: 128-38. http://dx.doi.org/10.12659/MSMBR.915067 PMID: 31040265

[62] Ma C, Wang F, Han B, *et al.* SALL1 functions as a tumor suppressor in breast cancer by regulating cancer cell senescence and metastasis through the NuRD complex. Mol Cancer 2018; 17(1): 78. http://dx.doi.org/10.1186/s12943-018-0824-y PMID: 29625565

[63] Li Z, Zhao S, Wang H, Zhang B, Zhang P. miR-4286 promotes prostate cancer progression *via* targeting the expression of SALL1. J Gene Med 2019; e3127. http://dx.doi.org/10.1002/jgm.3127 PMID: 31693770

[64] Gao Y, Ding Y, Chen H, Chen H, Zhou J. Targeting Krüppel-like factor 5 (KLF5) for cancer therapy. Curr Top Med Chem 2015; 15(8): 699-713. http://dx.doi.org/10.2174/1568026615666150302105052 PMID: 25732792

[65] Chen P, Qian XK, Zhang YF, Sun XG, Shi XJ, Gao YS. KLF5 promotes proliferation in gastric cancer *via* regulating p21 and CDK4. Eur Rev Med Pharmacol Sci 2020; 24(8): 4224-31. PMID: 32373958

[66] Wu Y, Qin J, Li F, *et al.* Usp3 promotes breast cancer cell proliferation by deubiquitinating klf5. JBC 2019; 294(47): 17837-47. http://dx.doi.org/10.1074/jbc.RA119.009102

[67] Guo C, Shi H, Shang Y, Zhang Y, Cui J, Yu H. LncRNA LINC00261 overexpression suppresses the growth and metastasis of lung cancer *via* regulating miR-1269a/FOXO1 axis. Cancer Cell Int 2020; 20(1): 275. http://dx.doi.org/10.1186/s12935-020-01332-6 PMID: 32607060

[68] Yan D, Liu W, Liu Y, Luo M. Linc00261 suppresses human colon cancer progression *via* sponging mir-324-3p and inactivating the wnt/β-catenin pathway. J Cell Physiol 2019; 234(12): 22648-56. http://dx.doi.org/10.1002/jcp.28831

[69] Liu S, Zheng Y, Zhang Y, *et al.* Methylation-mediated LINC00261 suppresses pancreatic cancer progression by epigenetically inhibiting c-Myc transcription. Theranostics 2020; 10(23): 10634-51. http://dx.doi.org/10.7150/thno.44278 PMID: 32929371

[70] Nikiforov YE, Nikiforova MN. Molecular genetics and diagnosis of thyroid cancer. Nat Rev Endocrinol 2011; 7(10): 569-80. http://dx.doi.org/10.1038/nrendo.2011.142 PMID: 21878896

[71] Corona RI, Seo J-H, Lin X, *et al.* Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. Nat Commun 2020; 11(1): 2020. http://dx.doi.org/10.1038/s41467-020-15951-0 PMID: 32332753

[72] Bie L-Y, Li N, Deng W-Y, Lu X-Y, Guo P, Luo S-X. Evaluation of PAX8 expression promotes the proliferation of stomach Cancer cells. BMC Mol Cell Biol 2019; 20(1): 61. http://dx.doi.org/10.1186/s12860-019-0245-9 PMID: 31881968

[73] Yokoyama T, Nakatake M, Kuwata T, *et al.* MEIS1-mediated transactivation of synaptotagmin-like 1 promotes CXCL12/CXCR4 signaling and leukemogenesis. J Clin Invest 2016; 126(5): 1664-78. http://dx.doi.org/10.1172/JCI81516 PMID: 27018596

[74] Ho JR, Chapeaublanc E, Kirkwood L, *et al.* Deregulation of Rab and Rab effector genes in bladder cancer. PLoS One 2012; 7(6): e39469. http://dx.doi.org/10.1371/journal.pone.0039469 PMID: 22724020

[75] Zhang M, Zhao J, Tang W, *et al.* High Hepsin expression predicts poor prognosis in Gastric Cancer. Sci Rep 2016; 6(1): 36902. http://dx.doi.org/10.1038/srep36902 PMID: 27841306

[76] Kim HJ, Han JH, Chang IH, Kim W, Myung SC. Variants in the HEPSIN gene are associated with susceptibility to prostate cancer. Prostate Cancer Prostatic Dis 2012; 15(4): 353-8. http://dx.doi.org/10.1038/pcan.2012.17 PMID: 22665141

[77] Willbold R, Wirth K, Martini T, Sültmann H, Bolenz C, Wittig R. Excess hepsin proteolytic activity limits oncogenic signaling and induces ER stress and autophagy in prostate cancer cells. Cell Death Dis 2019; 10(8): 601. http://dx.doi.org/10.1038/s41419-019-1830-8 PMID: 31399560

[78] Goel MM, Agrawal D, Natu SM, Goel A. Hepsin immunohistochemical expression in prostate cancer in relation to Gleason's grade and serum prostate specific antigen. Indian J Pathol Microbiol 2011; 54(3): 476-81. http://dx.doi.org/10.4103/0377-4929.85078 PMID: 21934206

[79] Nakamura S, Kanda M, Koike M, *et al.* Kcnj15 expression and malignant behavior of esophageal squamous cell carcinoma. Ann Surg Oncol 2020; 27(7): 1-10. http://dx.doi.org/10.1245/s10434-019-08189-8

[80] Liu Y, Wang H, Ni B, *et al.* Loss of *KCNJ15* expression promotes malignant phenotypes and correlates with poor prognosis in renal carcinoma. Cancer Manag Res 2019; 11: 1211-20. http://dx.doi.org/10.2147/CMAR.S184368 PMID: 30799948