Cancer Classification through a Hybrid Machine Learning Approach

Elmira Amiri Souri, Sophia Tsoka

Department of Informatics
Faculty of Natural and Mathematical Sciences
King's College London
London, WC2B 4BG, United Kingdom
Email: elmira.amiri@kcl.ac.uk, sophia.tsoka@kcl.ac.uk

***Abstract*—Understanding the underlying principles of cancer is a key endeavour in biomedical data mining. Although machine learning methods have been successful in discriminating normal from cancerous tissue with good accuracy, understanding of progression and formation of cancer across various cancer types is still restricted. Since cancer is a complex disease, being able to identify subgroups and investigate them separately may help in increasing the depth of our knowledge in terms of driver genes and oncogenic pathways. Moreover, as genes never act in isolation, methods that focus on single genes individually may be less efficient in uncovering key underlying molecular interactions. Algorithms that are capable of discovering the effect of combinations of genes have the potential to pave the way for extracting a new class of gene signatures that are neither mutated nor expressed differently, but rather act as mediators in forming oncogenic pathways. Here, we present a hybrid machine learning model to find cancer subgroups and an associated set of marker genes. In the proposed model, *autoencoders* are used to create a rich compressed set of features to identify cancer subgroups. Then, a two-step algorithm is developed based on information theory and regression analysis to find a set of discriminatory genes for each selected group for different types of cancer. This analysis is conducted based on the combined expression of genes to discover a new subset of genes associated with cancer. We show that we can still predict cancer accurately by decreasing the number of genes from thousands to tens for each subgroup. Pathway enrichment analysis is performed to find important pathways associated with a specific cancer type. The model is extensively analysed on datasets across nine cancer types and links between cancers are studied based on common gene signatures.**

Keywords–Machine Learning; Disease Classification; Clustering; Cancer Prediction.

I. INTRODUCTION

Cancer is a major cause of reduction in quality of life, with about 18.1 million new cases and 9.6 million cancer deaths noted recently (2018 [1]). Early detection of cancer can significantly improve prognosis, therefore, understanding the biology of cancer especially with regards to early detection is vital. Traditionally, clinical features such as age, tumor size, and cancer stage have been used to assist the prognosis of cancer, however these are only useful in late stage diagnosis and may not aid prediction [2].

High throughput technologies, such as microarray gene expression profiling and next-generation sequencing have produced an enormous amount of data which can be used to dissect cancer more accurately [3]. Early detection necessitates understanding the mechanism of cancer development via relevant associated and biological pathways. However, heterogeneity of tissues and genetics of patients prevent the identification of robust biomarkers [4] and the high dimensionality of expression data renders the selection of relevant genes in different types of cancer difficult [5]. Finally, as genes do not act in isolation and their combined effects lead to a variety of resultant phenotypes, the complexity of biomarker signatures increases [6].

Recently, machine learning and deep learning methods have resulted in advancement in the capability of prediction in many research fields with big and complex data, with notable applications in cancer research [7]–[10]. Deep learning methods have illustrated excellent potential in handling large and complex datasets and, together with the availablity of appropriate cancer profiling datasets [11], enable applications that can divulge key biomarker genes and pathways for disease types and increase our understanding of the mechanistic basis of cancer [8].

Identifying subgroups of similar pattern facilitates understanding of disease formation and progression. Once cancer subgroups are extracted, feature selection can be used for knowledge discovery through identification of key gene signatures [12]–[14]. Typically, methodologies rely on differentially expressed genes (for example, use of SAM [15], RVM [16] and SMVar [17]). However, these methods only focus on single genes and do not reflect the fact that genes work in functional groups. Additionally, there are genes contributing to cancer which may not be differentially expressed but may rather act as mediators in oncogenic pathways within a cancer network, establishing the connections between genes that are mutated or transcriptionally altered. Related work includes the work by Ghanat Bari et. al [9] that employ many concurrent Support Vector Machine models to derive a new class of cancer-related genes (named Class II genes) that are neither mutated nor differentially expressed, but proposed to act as potential key mediators in creating networks of cancer.

This work reports the development of a pipeline where the first stage involves application of an autoencoder, an unsupervised deep learning-based model, to compress high-dimensional gene expression data. Then, clustering is performed on the compressed gene expression data to discover different cancer subgroups, then each is assigned into two main classes called, *pure* and *mixed* based on the relevant sample

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4

label. Tumours which are very different from normal tissues form the *pure* groups, while tumours that are similar to normal samples fit into the *mixed* subgroups (mix of normal samples and tumours) . In the second stage, for each of these subgroups, a subset of gene biomarkers is selected through unsupervised (for *pure* subgroups) or supervised (for *mixed* subgroups) algorithms. The supervised method is a two-step algorithm based on information theory and regression analysis. Figure 1 shows the proposed framework. This approach was implemented for each of nine cancer types and it was shown that the derived gene markers are efficient in disease prediction. To highlight key cancer mechanistic details, pathway enrichment analysis was also applied and the network between different cancer based on common biomarkets was investigated. In Section II, the materials and methods applied in this paper are reviewed. Section III presents the results of the framework. Section IV concludes the paper and goes over the future work.

II. Materials and Methods

Gene expression data corresponding to nine cancer types were obtained from Gene Expression Omnibus (GEO) [18] for Affymetrix Human Genome U133 Plus 2.0 platform [9]. A total of 6957 cancer and 1850 normal tissue samples were collected. Table I shows the list of cancer data used in this paper. For pathway analysis, 188 KEGG [19] pathways were downloaded from GSEA, Broad Institute [20]. Raw Affymetrix data were normalised through Robust Multichip Average (RMA) [21] through the R BioConductor `rma` function [22]. Probes were mapped to genes by the Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip HG-U133_Plus_2) using the R Bioconductor annotation package `hgu133plus2.db` [23]. In cases of multiple probes mapping to the same gene, the average value of these probes is taken.

Datasets were split into training (90%) and external validation (test) set (10%) and on the training dataset all metrics were calculated through a 10-fold cross validation scheme, repeated 5 times. The training and test datasets are scaled by `StandardScaler`. To balance data, we applied Synthetic Minority Over-sampling Technique [24] using `SMOTE` function from imbalanced-learn 0.4.2 Python package to the training dataset to prevent overfitting on one class. Since the number of features (genes) is much larger than the number of samples, we should avoid to decrease the number of sample for balancing the data, therefore, oversampling is performed on training data.

To compress the expression of genes to the smallest set, autoencoder [25] was used. It is implemented using a multilayer neural network with a hidden layer in the middle and consists of two parts of encoding ($\phi : \chi \rightarrow F$) and decoding ($\psi : F \rightarrow \chi$). The loss function is defined in a way that the output is reconstructed from the input. Autoencoder is implemented by using Tensorflow 1.12.0 with three hidden layers and `Tanh` activation. Then, to identify groups of patients with similar gene expression patterns, several clustering algorithms were implemented (e.g., k-means, Spectral Clustering, Gaussian Mixture Models) in Scikit-learn 0.21.2 with default hyperparameters. As the successful method depends on the actual structure of the dataset [26], we found that for the size and nature of our data, k-means performed well (for an extensive study of clustering algorithms on large datasets, see [27]). For the implementation of clustering algorithm `MiniBatchKMeans` function with random initializations number= 3, batch size= 100, and reassignment ratio= 0.01 was used. The best number of clusters is selected by silhouette index [28]. After clustering, each sample is assigned to one of the modules; Modules with samples of the same label will be considered *pure*, whereas clusters with mixed labels (normal and tumor) will be identified as *mixed*.

$$C_i = \begin{cases} pure & \text{if } n_i^t/n_i^n < \alpha \\ & \text{or } n_i^n/n_i^t < \alpha, \\ mixed & \text{otherwise.} \end{cases}$$

where C_i is the i^{th} cluster, n_i^t and n_i^n are the number of tumor and normal samples in cluster i respectively, and α is a threshold set to 0.1.

The next stage involved finding a subset of biomarkers that can best characterise samples in each cluster. For *pure* clusters, since the label of all samples is the same, unsupervised feature selection was used, whereas in the case of *mixed* clusters, supervised feature selection was applied. Specifically, in the *pure* cluster, Principle Component Analysis (PCA) was applied to compress gene expression features and the overall contribution of each gene forming the principle components calculated by applying an inverse transform of the PCA to an identity matrix to observe which features had the highest contribution. For the implementation of the first step of our feature selection algorithm, `SelectKBest` function with `mutual_info_classif` score function and for the second step `LassoCV` were used. Similarly, to perform PCA, we used `PCA`. In the case of *mixed* clusters, a two-step feature selection algorithm called *BestLasso* was implemented based on combination of information theory and LASSO (Least Absolute Shrinkage and Selection Operator) [29]. In BestLasso algorithm, first a subset of the highest contributing features is chosen by estimating the mutual information [30] of every feature with the labels, then Lasso was used to select the best set of features. The main reason for performing this two-step process is because gene expression data has high dimensionality and performing Lasso on all data becomes prohibitively slow and complex. Algorithm 1 shows this procedure.

Differentially Expressed (DE) genes were selected by calculating t-test in R `Limma 3.26.9` package. The p-value was adjusted by the moderated t-test for multiple testing by `BH-adjusted` (Benjamini-Hochberg method). We used `topTable` function from limma with log-fold-change `(logFC)` $> |2|$. [-23pt]

Once features were selected, classification was performed by learning a model on the selected features to predict

TABLE I. The list of cancer data used in this paper

Cancer	# of samples	# of tumor samples	# of normal samples
Breast	2113	1984	129
Ovary	954	839	115
Colon	1765	1557	208
Prostate	389	299	90
Skin	621	357	264
Liver	588	279	309
Pancreatic	259	178	81
Kidney	1031	589	442
Lung	1087	875	212
Total	**8807**	**6957**	**1850**

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4

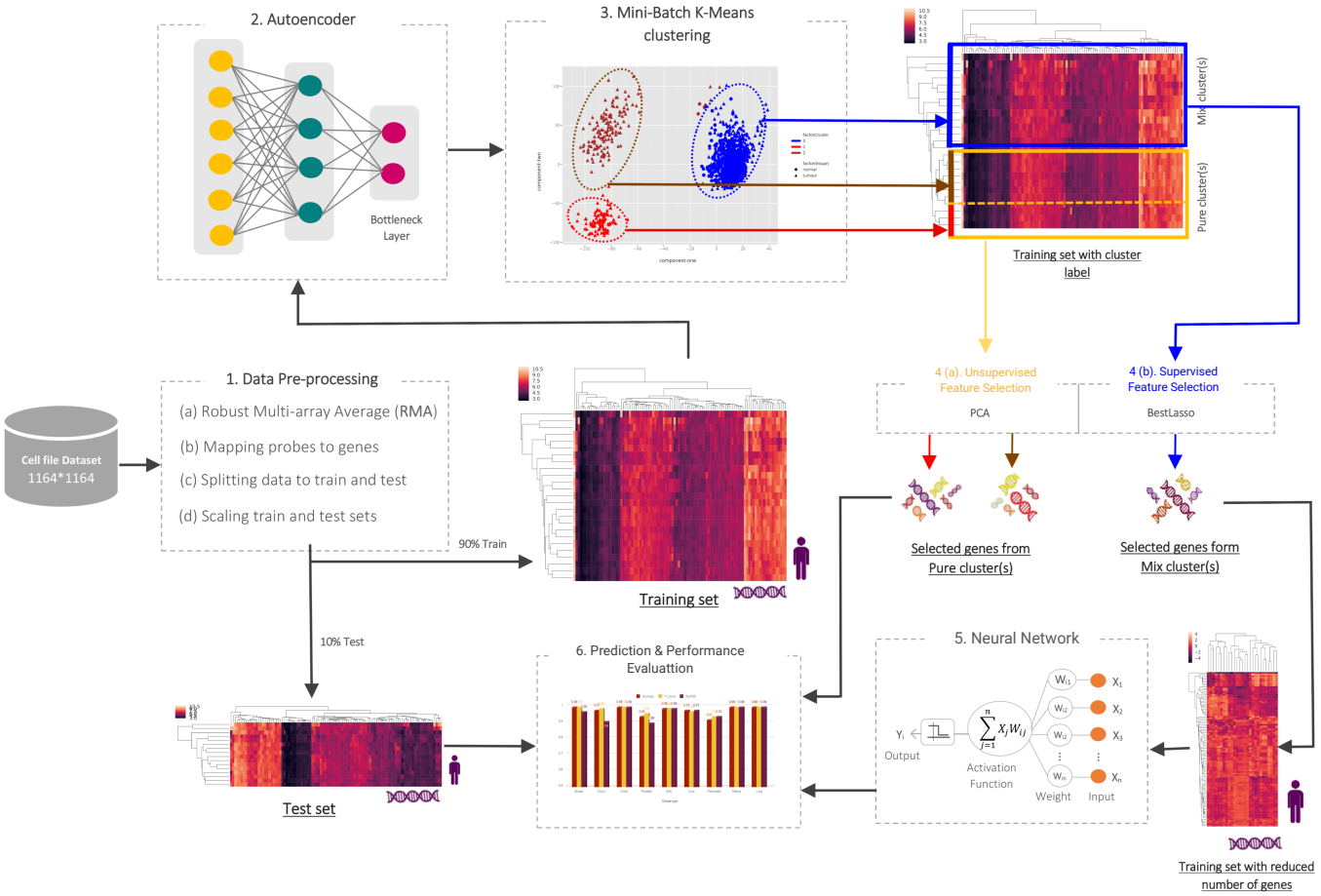

Figure 1. Overview of the proposed model to infer cancer subtypes and gene signatures (data here relate to the breast cancer dataset). 1.Preprocessing data. 2. An Autoencoder applied to compress the high-dimensional set of genes 3. The samples are grouped into different subgroups based on the encoded features. The cluster labels are added to each sample 4(a). All the genes of the subgroups that have same tissue types (*pure* clusters) are fed to PCA to select their highest contributing genes. 4(b). All the genes of the subgroups with *mixed* tissue types along with their labels are then input into a feature selection algorithm, BestLasso. 5. The selected features are used in a neural network to learn a model able to predict their labels. 6. The whole model is evaluated on the test set 5 times 10-fold cross-validation.

Algorithm 1 BestLasso algorithm. $X_{N\times m}$ is the gene expression data of N samples and m genes. Y is the label for each sample (0 and 1). k is the number of desired features from Mutual Information algorithm. $\mu(f_i)$ is the probability density for sampling f_i.

procedure BESTLASSO($X_{N\times m}, Y, k$)
  $feature\_scores \leftarrow [\,]$
  for $i = 1 \rightarrow m$ **do**
    $f_i \leftarrow X[:, i]$
    $s_{ia} \leftarrow \sum_{Y} \int \mu(Y, f_i) \log \mu(f_i|Y) df_i$
    $s_{ib} \leftarrow \int \mu(f_i) \log \mu(f_i) df_i$
    $s_i \leftarrow (s_{ia} - s_{ib})$
    $feature\_scores[i] \leftarrow s_i$
  $X'_{N\times k} \leftarrow$ Top k features with highest $feature\_scores$
  return $Lasso(X', Y)$

tissue type (normal or tumor) in *mixed* clusters as samples in *pure* clusters. Different techniques were evaluated, including Logistic Regression, Support Vector Machine and Random Forrest, and neural networks in Scikit-learn 0.21.2 with default hyperparameters. Among them neural networks had roughly better average performance on 9 cancer datasets. Then, neural networks hayperparameters were tuned using `GridSearchCV` and a model with an input layer, two hidden layers (30 and 5 nodes, 'relu' activation function) and an output layer with 'sigmoid' activation function was chosen. Five times 10-fold cross-validation was done on the training data and the test set data was used for evaluation of the classification procedure through accuracy, F1 score, and area under ROC curve (AUROC) metrics.

III. RESULTS AND DISCUSSION

Cancer is a heterogeneous disease with different histopathological and molecular subtypes, each with different diagnosis and therapies [31]. The goal of this work is

to propose a way to find these subtypes by maximising the intra-group and minimising inter-group similarity [32]. Sample stratification is difficult when relying only on clinical data, therefore the use of gene expression facilitates more meaningful pattern extraction and sample stratification. To achieve this target, machine learning clustering methods can identify groups of patients with similar gene expression patterns. In this section, we present the results of our method in selecting *pure* and *mixed* subgroups on nine different cancer types. We then introduce the genes related to cancer for each subgroup and report the high performance of the model using these genes. Lastly, we conduct pathway enrichment analysis and show that some of these results can provide validation through existing relevant literature.

In order to discover subgroups, a key step in clustering is determining the optimal number of clusters. A common method to perform this is by evaluating clusters using cluster validity indices, where samples are closely linked within the same cluster and are well-separated from members of other clusters [28]. To identify the optimal number of cancer subgroups, the average Silhouette score of 5 different number of clusters was computed. For breast cancer for example, the best way of subgrouping data is with three clusters (Figure 1 section 3). One of these subgroups is a *mix* of normal and tumor tissues, while the other two contain only tumor samples i.e., *pure* cluster. For all cancers, Table II shows the average Silhouette scores for each type respectively. In most cases, clustering has been able to separate the samples well using the compressed set of genes.

TABLE II. DIFFERENT SIZES OF CLUSTERS AND THEIR AVERAGE SILHOUETTE SCORE

Cancer \ Number of Clusters	2	3	4	5	6
Breast	0.80	**0.82**	0.60	0.55	0.49
Ovary	0.68	**0.73**	0.46	0.47	0.36
Colon	**0.73**	0.55	0.42	0.41	0.38
Prostate	0.59	**0.70**	0.55	0.60	0.54
Skin	0.53	0.56	**0.61**	0.58	0.51
Liver	0.41	**0.52**	0.50	0.39	0.39
Pancreatic	0.55	**0.62**	0.58	0.48	0.42
Kidney	**0.69**	0.50	0.48	0.44	0.41
Lung	**0.51**	0.44	0.34	0.33	0.35

After the optimal subgroups and the type of clusters (*pure* or *mixed*) are identified, gene signature subset selection was performed to extract useful information in each subgroup and reduce dimensionality (out of more than 22,000 genes). Since each cluster represents a different cancer subgroup, studying the selected genes in each cluster individually will lead to the identification of relevant gene signatures. One of the common methods of ranking genes associated to cancer is by selecting genes expressed differently in tumor and normal tissue using statistical methods. Selecting only Differentially Expressed (DE) genes results in genes being considered individually, regardless of their inter-relationships. As traits and phenotypes are caused by interactions of groups of genes [33], here we use a powerful machine learning strategy that can test for different combination of genes sets as means for deriving robust cancer biomarkers that have the ability of predicting cancer with high accuracy. Table III contains the list cancer subgroups and the number of their gene signatures. For example, breast cancer consist of two *pure* and one *mixed* subgroups with different number biomarkers selected in each. A list of the biomarkers for each cancer subgroup is given in Table IV.

TABLE III. LIST OF OPTIMAL SUBGROUP TYPES AND NUMBER OF GENE SIGNATURES IN EACH OF THEM

Cancer	Cluster Types (# of Gene Signature)			
Breast	Pure (28)	Pure (36)	Mixed(70)	
Ovary	Pure (41)	Mixed (90)	Mixed(19)	
Colon	Pure (normal)	Mixed (82)		
Prostate	Mixed (29)	Mixed(53)	Mixed(25)	
Skin	Pure (normal)	Pure (49)	Pure(38)	Mixed(28)
Liver	Mixed (34)	Mixed (27)	Mixed (45)	
Pancreatic	Pure (27)	Mixed (26)	Mixed (9)	
Kidney	Pure (normal)	Mixed (59)		
Lung	Pure (30)	Mixed (56)		

Moreover, methods that rely on just differentially expressed genes ignore mediator genes which are contributing to cancer but may not mutated. Recently, methods that aim to delineate such genes active in connecting oncogenic pathways are reported [9]. From all the gene signatures selected by our framework, some of them are differentially expressed and some are not, which may indicate mediator genes. As an example, mediator genes in breast cancer found by our model are as follows: *ABCA8*, *ARCN1* [34], *ARHGAP20* [35], *ATP5B* [36], *CA4* [37], *CLDN5* [38], *DCTN2*, *FAM13A*, *GLYAT*, *GRIP2*, *GSTM5* [39], *H3F3A*, *HIST1H3I*, *KIF23* [40], *NUP210* [41], *RAB7A* [42], *RPL7A* [42], *RPLP0*, *RPS12*, *SIN3A* [43], *SPTBN1*, *TUBA1C* [44].

Since there are multiple gene signatures common between each cancer, a network of cancers can be outlined. Figure 2 shows this network comprised of all the chosen gene signatures by our model colored based on the 9 different cancer types. Each cancer has their own gene signature while some of them share specific genes, as indicated in the figure. Our analysis showed that *ABCA8* is a hub gene shared between four cancers and known to be involved in multiple cancers in literature [45]–[47]. Another interesting observation is the many common gene signatures between breast and lung cancers: *CA4* [37], *FIGF* [48] , *LDB2* [49], *GPIHBP1* [50], *COL10A1* [51], *SLC19A3* [52], *LYVE1* [53], *IGSF10* [54] , *MYZAP* [55], *SPTBN1*, *ADH1B* [56], *ABCA6* [57], *PIR-FIGF*. Almost all of them are also reported as being associated with lung and breast cancer. It is note that lung is the most likely tissue for cancer metastasis from breast [58] [59].

The results of the prediction of the proposed model on the test set are presented in Table V. The model is performing with higher than 90% accuracy and F1-score in all cases which means that the set of selected genes are capable of accurately distinguishing between cancer and normal tissues. The two cancers with the lowest accuracy are Prostate and Pancreatic cancers, for which the lowest number of samples was available. It is noted that the model may improve upon availability of a larger data size for these cancers.

Once all important genes are selected and validated by our method, we can gain further insight through pathway enrichment analysis for each subgroup. To this end, the number of selected gene signatures in each pathway is determined and normalised by the total number of genes in the pathway, the counts therefore serving to demonstrate the importance of the pathway in the cancer subgroup. Some key pathways are already known as pathways associated with cancer and some of them have not been studied specifically yet and can be aimed for further research. Full list of the most important pathways

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4

TABLE IV. LIST OF BIOMARKERS FOR EACH SUBGROUP

Cancer Subtype	Biomarkers
breast (mixed)	ABCA6, ABCA8, ADAMTS5, ADH1B, ADH1C, ALDH1L1, ANXA1, ARHGAP20, ARID5B, ATOH8, C2orf40, CA4, CD300LG, CEP55, CLDN5, CLEC3B, CNRIP1, COL10A1, COL11A1, COPG2IT1, DPT, FAM13A, FIGF GINS1, GJB2, GLYAT, GPIHBP1, GSTM5, HELLS, HSPB2, HSPB7, IGFBP6, IGSF10, INHBA, ITIH5, KIF14, KIF23, KLF15, LDB2, LINC01614, LRRN4CL, LYVE1, MAMDC2, MATN2, MME, MYZAP, NPR1, NUP210, PAFAH1B3, PAMR1, PGM5, PLAC9, PLIN1, RGN, RRM2, SBK1, SCARA5, SCN4B, SIK2, SLC19A3, SMC4, SPATS2, SPTBN1, TMEM246, TNMD, TNXA, TRIM59, TSHZ2, UHRF1, VIT
breast (pure)	SNHG7, LOC283674, RPS6KA2-AS1, C9orf50, RPL13A, TSPAN16, FLJ31713, RPS12, CFAP100, LINC00967, RPL7A, LOC101928602, LOC100288123, XKR7, HPR, LOC101929738, LOC101929144, FCAR, ACTG1, ZCCHC13, ARCN1, RPLP0, LOC101929680, TUBA1A, TUBA1C, ATP5B, TMEM203, SNORA74A
breast (pure)	HIST1H3I, RFFL, LOC100505716, GRIP2, SLC6A17, LOC645513, RBM26, NENF, C5orf51, APMAP, MLLT10, DHRS7, HDGFL1, IL10RB-AS1, LDLRAD4-AS1, SIN3A, PRDM2, LOC100506858, FKSG29, DANT2, LOC105370977, CACNG6, RAB7A, TMEM161B-AS1, LRRC43, EMC7, DCTN2, USF3, H3F3A, TRAFD1, LOC84843, MTPN, LINC00641, REST, TH2LCRR, RNF152
colon (mixed)	ABCA8, ABCG2, ADAMDEC1, ADH1C, ADTRP, AJUBA, AMPD1, APPL2, BEST4, C15orf48, C2orf88, CA1, CA2, CA7, CDH3, CDKN2B, CEMIP, CHGA, CHP2, CLCA4, CLDN1, CLDN23, COL11A1, CSE1L, CWH43, DHRS11, DUSP14, EDN3, ENTPD5, ETHE1, FKBP1A-SDCBP2, FLJ36848, FOXQ1, FUCA1, GCG, GPAT3, GPD1L, GTF2IRD1, GUCA2B, HAPLN1, HIGD1A, HILPDA, HPGD, INHBA, ITM2C, KIAA1549, KLF4, KRT80, LIFR, LINC00675, LPAR1, LRRC19, MOGAT2, MRGBP, MTHFD1L, NAAA, NFE2L3, NR3C2, P2RX4, PDCD4, PLCL2, PLP1, PRDX6, PYY, SCARA5, SLC25A34, SLC51B, SLC6A6, SLC7A5, SMPDL3A, SNTB1, SPPL2A, SST, TEAD4, TMCC3, TPH1, TRIB3, TSPAN1, TSPAN7, UGDH, VSTM2A, ZG16
kidney (mixed)	ABAT, ACOX2, ALAD, APEH, AQP2, ASS1, ATP6V0D2, CA9, CALB1, CAPN3, CLCNKB, CLDN10, COL23A1, CRYAA , CTSH, DCXR, EFCAB3 , EGLN3, ENPP6, ERP27, FBP1, FBXO16 , FOXI1, FXYD4, GATA3, GGH, HRG, HS6ST2, IGFBP3, IRX1, KCNJ1, KLHL13, KLHL14, KNG1, LARS2, LINC00887, LOC100130278, LOC101928574, LOC102723468, MT1G, NOL3, NPHS2, OAT, PTH1R, RDH11, RGS1, S100A2, SCNN1G, SERPINA5, SLC12A1, SLC25A5, TFAP2B, TMEM213, TMEM30B, TMEM52B, TMPRSS2, TNFAIP6, VIM, ZNF395
liver (ixed)	ADAMTS13, ANXA3, BEX1, BIRC5, BMP5, CFP, CNDP1, COL15A1, CYB5D1, CYP2C8, DACH1, DBH, DCUN1D3, DPF3, F9, GPM6A, HHIP, HSPB1, ITLN1, KAZN, KIAA0907, LCAT, LHX2, LINC01296, MAP2K1, MT1G, MYOM2, NSUN5, NSUN5P1, OLFML2B, PLAC8, PLVAP, POGZ, PROM1, PTH1R, SLC16A5, SLC46A3, SLC5A1, SLCO4C1, SNX27, STAB2, TARBP1, TCF21, THY1, WDR66
liver(mixed)	ADGRG7, ADK, ANGPTL3, BLOC1S1-RDH5, C1orf168, CAP2, CENPF, COL25A1, DGAT2, EPS8L3, ESR1, FREM2, KCNJ16, LAMC1, MT1H, NAPSB, PAMR1, PEG3-AS1, PLCB1, PPM1H, RANBP3L, RPS6KA6, SESTD1, SHC1, SSR2, STEAP3, TREH
lung (mixed)	ABCC3, ADCY4, ADRA1A, AGR2, AKAP2, AMOTL1, ARHGAP6, ASPRV1, BVES, CA4, CCBE1, CDH5, COL10A1, DACH1, FGD5, FGFR4, FOXF1, FUT2, GCNT3, GPRC5A, GRK5, HABP2, HSH2D, IGSF10, KDELR3, LIN7A, MAGI2-AS3, MUC20, MYCT1, NCKAP5, P2RY1, PAK1, PEAR1, PHF2, PPM1F, PROM2, RASIP1, RHBDL2, SDC1, SEMA6A, SGCG, SH3GL2, SH3GL3, SLC19A3, SLC39A11, SOX17, SPINK1, SPOCK2, SPTBN1, STARD13, TAL1, TGFBR3, TMPRSS4, TSPAN18, WFDC2, ZBED2
lung (pure)	LDB2, LYVE1, SDPR, ABCA6, FAM150B, ARHGAP6, RHOJ, AGER, ADH1B, EMCN, GPIHBP1, MMRN1, GRK5, GPM6A, MYZAP, ABCA8, S1PR1, FIGF, ASPA, ANGPTL1, NME1, GRIA1, CA4, EDNRB, PTPRB, SCN7A, TCF21, PCAT19, TEK, FHL1
ovary (pure)	P4HB, NOP10, SRP9, FTL, CDC37, MIF, NBPF10, CHMP2A, ARF1, COPZ1, MRPL37, NDUFAB1, SCAND1, RHOA, PGRMC1, XRN2, PSMC3, POLR2E, EIF4A1, DDOST, SPCS2 , GNB2, TUBA1C, ABHD17A, PRPF31, NDUFA3, PCBP1, RPS27, OST4, OAZ1, APEX1, UBC, RNF181, JTB, TMEM258, RPS5, MRPL34, HSPA8, H3F3A, CHCHD2, LSM7
ovary(mixed)	ABCA8, ABHD11, ABHD17C, ADGRD1, ANKRD29, AOX1, AP1M2, ARHGAP8, ARMCX5-GPRASP2, ARX, ASS1, ATP10D, BAMBI, BDH2, C14orf37, C1orf186, CACNB2, CD24, CELF2, CHD7, CLDN4, CLDN7, CNIH3, CNRIP1, CP, CPED1, CSGALNACT1, CXXC5, CXorf57, DFNA5, ECM2, EPCAM, FAM153A, FAM153A , FLRT2, GHR, GNG11, GPRASP1, GRHL2, HAND2-AS1, HOXC6, IDH2, KCNT2, KLHL14, KPNA5, L3MBTL3, LEMD2, LIN7A, LOC728392, MAF, ME1, MECOM, MUC1, MUM1L1, NBEA, NDNF, NR3C2, OLFML1, PEG3, PID1, PLCL2, POLR3GL, PPM1K, PPP4R4, PRSS35, RNASE4, RPL36A, SERP2, SIGLEC11, SLC30A4, SLC34A2, SLC44A2, SNCA, SORT1, SPINT1, STON2, SYTL1, TCEAL2, TCEAL3, TCEAL7, TES, TFPI, TLE4, TMEM139, TMEM150C, TRIM68, TRPC1, TSPAN5, WFDC2, WHAMMP2
ovary (mixed)	ARID4B, CASP2, FMN2, GS1-259H13.2, HIST1H3I, HPS3, KLHL24, NCOA2, NICN1, PCED1B, PLXND1, PPIAP21, PSMG3-AS1, RAB4A, SHROOM2, TOPBP1, TUBB4B, VSIG1, ZDHHC20
pancreatic (mixed)	FBXO25, HOXC6, NRG4, PDIA2, PRR11, RPL14, SLC25A13, SND1, TNFAIP1
pancreatic (mixed)	AFAP1-AS1, ARMC9, CALU, CLDN1, CLDN4, CST1, CTTN, HIST2H2AA3, HOXB7, HOXC6, KRT18, LOC340340, MAMDC2, MROH6, MSLN, NAT14, NME1-NME2, PKM, PLXNB2, PYGB, RPL23A, SDC1, SDC4, SLPI, TTTY5, VGLL4
pancreatic (pure)	CPEB2, SNHG10, LOC642862, COQ10B, AFF4, HIST1H4B, C6orf106, ARID5B, CDK9, LOC100129112, CCDC117, BOLA2, NOCT, POLR2A, PRDM2, ZFX, C16orf72, B4GALT1, GATAD2A, ATXN2L, LOC101926943, AHNAK, CCNK, RAB7A, CDR1, MTPN, ZNF460
prostate (mixed)	ACSS2, CDKN2A, CPSF7, ENTPD3, FADS1, FAP, IGSF1, KANK4, LINC00328, LINC00869, LOC100996741, LOC158863, LOC441666, NETO2, NFAT5, PCSK5, RNF24, SALL3, SMIM10L2A, SPPL3, ST3GAL5, TMCO3, TMEM241, ZNF595, ZNF93
prostate (mixed)	ACOX2, AMACR, CFC1, COL9A1, CYP4B1, EFS, FHL2, FLRT3, FOXQ1, HADHB, LSAMP, MME, MSMB, MSMO1, NEFH, NPM1, PCAT4, RBBP7, SMIM5, WIF1
prostate (mixed)	ADCY4, ADORA2A, ADRB1, ARHGEF15, ARRDC2, ATHL1, ATP7A, BCKDHB, C10orf10, C1R, CADM3, CHST7, CLEC14A, CLIC2, CNBD2, DMBT1, DOCK9, FAM193B, FECH, FES, FHL5, GIMAP1, IGFBP5, IL15RA, KCNMB2, LCLAT1, LGR4, LINC01503, LOC100507291, LOC100996583, LOC10537679, LOC286071, LYPLA1, MAP3K3, MFAP3, MS4A14, MSC-AS1, NPR2, PDGFRA, PDLIM1, PNPLA4, PPP6R1, PSMA5, SHC1, SLC39A9, TIE1, TMEM218, TMEM255B, TMOD1, TRIP10, TSC22D1, TTR, UGP2
skin (pure)	SMAD1, SLC46A2, CCDC186, NIPAL1, DENND4C, XG, NET1, MYO6, HLF, ATP8B1, THRB, FOXN3, BCL11B, GIPC2, RAPGEFL1, ABHD5, LNX1, CEBPG, MAF, LRBA, LOC284023, RORA, TMTC3, CCDC6, TTC39B, GLTP, DENND2C, MPZL3, F3, PPM1L, ABLIM1, ELOVL4, FBXW7, TUFT1, GAN, ACVR2A, ELL3, LOC101927164
skin (pure)	TBC1D8, LOC102724593, THRA, TMEM262, SIPA1L3, MMP19, XGY2, RPARP-AS1, LOC100132319, SPAG8, ELMSAN1, ESRG, SPIDR, CYP4Z1, PCNT, ADIRF-AS1, LOC101928988, IL17RE, NUDT17, CCDC153, SAPCD1-AS1, LOC283713, EEF1D, LIPH, YPEL2, CDR1, MIR4697HG, DCST2, RPRML, LOC105369671, UBE2NL, SLC9A3R2, AGAP11, ANKRD19P, CENPT, TYSND1, AP1G2, RRN3P3, HSPA1B, LOC101928595, LOC105375061, LOC105379661, SKIDA1, ACTA2-AS1, LOC102723600, GATB, RNF31, FOXH1, CYP21A1P
skin (mixed)	ADAM12, APOBEC3C, ARPC1B, ATL1, BCL2A1, BMP2, BTC, CLDN23, CLDN8, CP, EPB41L4B, FCMR, HN1, HPGDS, IFI27, IGFL1, ITPA, LOC105378074, MIR503HG, MSH5-SAPCD1, NDC80, OASL, PIK3CD, PRDM6, RTP4, SGCG, SLC8A1, TMEM206

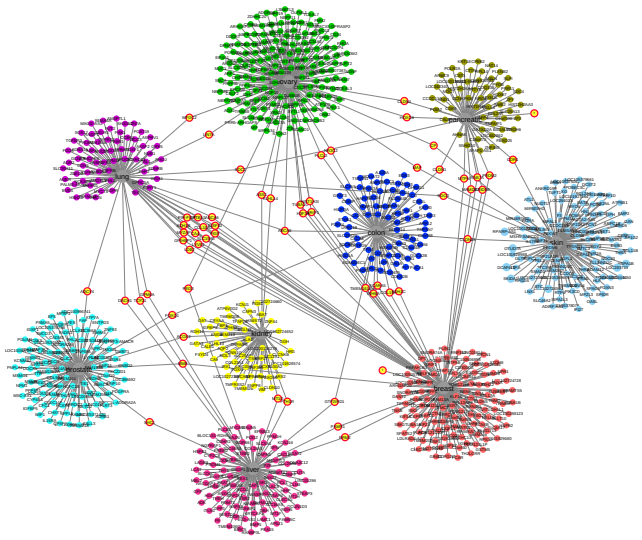

Figure 2. Gene signatures selected by our framework coloured according to cancer type. The common genes connecting more than one cancer are shown in orange.

TABLE V. THE ACCURACY, F1 SCORE, AUROC OF OUR MODEL ON 9 CANCER TYPES

Cancer	Acurracy	$f1_{score}$	AUROC
Breast	0.99	0.99	0.96
Ovary	0.97	0.98	0.90
Colon	0.99	0.99	0.99
Prostate	0.93	0.98	0.89
Skin	0.98	0.98	0.98
Liver	0.97	0.96	0.97
Pancreatic	0.91	0.93	0.93
Kidney	0.99	0.99	0.99
Lung	0.99	0.99	0.99

among KEGG pathways for each subgroup summarized in Table VI. Moreover, from our analysis, there are several pathways repeated in multiple cancer subgroups and all these pathways are cited in the literature as associated to cancer. Table VII shows key pathways identified through the KEGG.

IV. CONCLUSION AND FUTURE WORK

With the advent of massively parallel profiling of genes and their products, as well as improved machine learning technologies to handle large and heterogeneous datasets, enhancing analyses for cancer is possible. In this work, we present a hybrid machine learning computational procedure that includes analysis of datasets from multiple cancer types, integration of supervised and unsupervised learning procedures in the same computational framework and the use of autoencoder step that can effectively compress the high dimensionality of the gene expression profiles to discovering cancer subgoups.

Using this approach we identified a set of genes involved in cancer, some of them being recently reported in literature. As another means of validation, we were able to perform classification with very high accuracy using these biomarkers on the test set. In addition to being able to accurately predict cancer, our goal was to increase understanding of the underlying mechanisms by performing analysis on the selected genes and their pathways. Therefore, a network was created to show common biomarker genes among different types of cancer, that can reveal relationships between cancer types, e.g., breast and lung, as previously noted. Additionally, pathway enrichment analysis on our data identified the most important KEGG pathways, with some of them known to have a role in cancer formation and progression. Finally, differentially expressed genes were computed and compared with the selected genes to identify a new set of genes that are believed to act as mediators. The suggested pipeline for subgrouping cancer represents a novel contribution towards analysing transcriptomic cancer tissue data and aiding the development of sophisticated machine learning methods for big, complex and noisy data. In future work, clinical aspects of each subgroup can be taken into consideration by including

TABLE VI. LIST OF IMPORTANT KEGG PATHWAYS FOR EACH SUBGROUPS OF ALL THE NINE CANCERS.

Cancer Subgroup	**KEGG Pathway**
breast (mixed)	RENIN_ANGIOTENSIN_SYSTEM ONE_CARBON_POOL_BY_FOLATE
breast (pure)	PATHOGENIC_ESCHERICHIA_COLI_INFECTION
breast (pure)	VASOPRESSIN_REGULATED_WATER_REABSORPTION
colon (mixed)	NITROGEN_METABOLISM
kidney (mixed)	VALINE_LEUCINE_AND_ISOLEUCINE BIOSYNTHESIS_FOLATE_BIOSYNTHESIS
liver (mixed)	DORSO_VENTRAL_AXIS_FORMATION
liver (mixed)	BETA_ALANINE_METABOLISM
liver (mixed)	RETINOL_METABOLISM
lung (mixed)	GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES
lung (pure)	ABC_TRANSPORTERS
ovary (mixed)	GLYCOSAMINOGLYCAN_BIOSYNTHESIS _CHONDROITIN_SULFATE
ovary (mixed)	DORSO_VENTRAL_AXIS_FORMATION
ovary (pure)	PROTEIN_EXPORT
pancreatic(mixed)	PATHOGENIC_ESCHERICHIA_COLI_INFECTION
pancreatic (mixed)	ERBB_SIGNALING_PATHWAY
pancreatic (pure)	GLYCOSAMINOGLYCAN_BIOSYNTHESIS _KERATAN_SULFATE
prostate (mixed)	PRIMARY_BILE_ACID_BIOSYNTHESIS
prostate (mixed)	GLYCOSAMINOGLYCAN_BIOSYNTHESIS _CHONDROITIN_SULFATE
prostate (mixed)	GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO _SERIES
skin (pure)	THYROID_CANCER
skin (pure)	RNA_POLYMERASE
skin (mixed)	ARRHYTHMOGENIC_RIGHT_VENTRICULAR _CARDIOMYOPATHY_ARVC

TABLE VII. LIST OF IMPORTANT KEGG PATHWAYS IN NINE CANCERS

Pathway	**Reference(s)**
PURINE_METABOLISM	Purines play a critical role in cell proliferation and their broken metabolism has recently been recognized to be related to cancer progression [60]
PATHWAYS_IN_CANCER	KEGG has identified a pathway which is related to cancer [19]
LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	Leukocytes cells are exploited by tumour cells for extravasation [61]
PYRIMIDINE_METABOLISM	Edwards et al. [62] have extensively studied human skin cutaneous melanoma (SKCM) and found pyrimidine metabolism as a major pathway in its progression.
MAPK_SIGNALING_PATHWAY	The role of mitogen-activated protein kinase (MAPK) pathways in cancer is studied in [63]. Changes in MAPK pathways can mainly affect Ras and B-Raf in extracellular signal-regulated kinase pathway.
FOCAL_ADHESION	Focal adhesion kinase (FAK) plays an important role in tumor progression and metastasis because it is in charge of cancer cell signalling, cell proliferation, cell survival and cell migration [64].
NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	He et al. [65] studied the gene expression in prostate cancer and found the neuroactive ligand-receptor interaction as one of the enriched pathways.

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4

them as relevant features, using them as prediction outcomes or validating biomarkers against them (e.g., use of survival data for validation). The desired outcome will be to enhance accurate cancer diagnosis, while also paving the way for evaluating therapeutic interventions.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," CA: a cancer journal for clinicians, vol. 69, no. 1, 2019, pp. 7–34.

[2] C. A. Borrebaeck, "Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer," Nature Reviews Cancer, vol. 17, no. 3, 2017, p. 199.

[3] L. De Cecco, P. Bossi, L. Locati, S. Canevari, and L. Licitra, "Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data defines a robust survival predictor," Annals of oncology, vol. 25, no. 8, 2014, pp. 1628–1635.

[4] S. Turajlic, A. Sottoriva, T. Graham, and C. Swanton, "Resolving genetic heterogeneity in cancer," Nature Reviews Genetics, vol. 20, no. 7, 2019, pp. 404–416.

[5] J. Li et al., "Identification of high-quality cancer prognostic markers and metastasis network modules," Nature Communications, vol. 1, Jul 2010, pp. 34 EP –, article.

[6] S. Gao et al., "Identification and construction of combinatory cancer hallmark–based gene signature sets to predict recurrence and chemotherapy benefit in stage ii colorectal cancer," JAMA oncology, vol. 2, no. 1, 2016, pp. 37–45.

[7] A. Penson et al., "Development of genome-derived tumor type prediction to inform clinical cancer care," JAMA oncology, vol. 6, no. 1, 2020, pp. 84–91.

[8] A. Rahimi and M. Gonen, "Discriminating early- and late-stage cancers using multiple kernel learning on gene sets," Bioinformatics, vol. 34, no. 13, 2018, pp. i412–i421.

[9] M. Ghanat Bari, C. Y. Ung, C. Zhang, S. Zhu, and H. Li, "Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks," Sci Rep, vol. 7, no. 1, Aug 2017, pp. 6993–6993.

[10] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning–based multi-omics integration robustly predicts survival in liver cancer," Clinical Cancer Research, vol. 24, no. 6, 2018, pp. 1248–1259.

[11] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," Computer Methods and Programs in Biomedicine, vol. 153, 2018, pp. 1 – 9.

[12] S.-B. Cho and H.-H. Won, "Machine learning in dna microarray analysis for cancer classification," in Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003 - Volume 19, ser. APBC '03. Darlinghurst, Australia: Australian Computer Society, Inc., 2003, pp. 189–198.

[13] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, 2007, pp. 2507–2517.

[14] J. Li and E. Wang, "A multiple survival screening algorithm (mss) for identifying high-quality cancer prognostic markers," Feb 2011.

[15] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," Proceedings of the National Academy of Sciences, vol. 98, no. 9, 2001, pp. 5116–5121.

[16] G. W. Wright and R. M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments," Bioinformatics, vol. 19, no. 18, 2003, pp. 2448–2455.

[17] F. Jaffrezic, G. Marot, S. Degrelle, I. Hue, and J.-L. Foulley, "A structural mixed model for variances in differential gene expression studies," vol. 89, no. 1, 2007, pp. 19–25.

[18] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," Nucleic acids research, vol. 30, no. 1, 2002, pp. 207–210.

[19] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," Nucleic Acids Res, vol. 28, no. 1, Jan 2000, pp. 27–30, 10592173[pmid].

[20] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," Proceedings of the National Academy of Sciences, vol. 102, no. 43, 2005, pp. 15 545–15 550. [Online]. Available: https://www.pnas.org/content/102/43/15545

[21] R. Irizarry et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," Biostatistics, vol. 4, no. 2, 2003, pp. 249–264.

[22] 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, 2004, pp. 307–315.

[23] M. Carlson, hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2), r package version 3.2.2.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," J. Artif. Int. Res., vol. 16, no. 1, Jun. 2002, pp. 321–357.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, 2006, pp. 504–507.

[26] P. D'haeseleer, "How does gene expression cluster work?" Nature biotechnology, vol. 23, 01 2006, pp. 1499–501.

[27] M. C. P. de Souto, I. G. Costa, D. S. A. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study." BMC Bioinformatics, vol. 9, 2008.

[28] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, no. 1, Nov. 1987, pp. 53–65.

[29] V. Fonti, "Feature selection using lasso," Research paper in business analytics, 2017.

[30] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E, vol. 69, Jun 2004, p. 066138.

[31] Z. Liu, X.-S. Zhang, and S. Zhang, "Breast tumor subgroups reveal diverse clinical prognostic power," Sci Rep, vol. 4, Feb 2014, pp. 4002–4002.

[32] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," IEEE Trans. on Knowl. and Data Eng., vol. 16, no. 11, Nov. 2004, pp. 1370–1386.

[33] D. Moore, The Dependent Gene: The Fallacy of "Nature vs. Nurture", 01 2002.

[34] J. T.-H. Chang, F. Wang, W. Chapin, and R. S. Huang, "Identification of micrornas as breast cancer prognosis markers through the cancer genome atlas," PLoS One, vol. 11, no. 12, Dec 2016, pp. e0 168 284–e0 168 284, 27959953[pmid].

[35] D. Oliver et al., "Identification of novel cancer therapeutic targets using a designed and pooled shrna library screen," Sci Rep, vol. 7, Feb 2017, pp. 43 023–43 023, 28223711[pmid].

[36] J. Cuezva et al., "The bioenergetic signature of cancer," Cancer Research, vol. 62, no. 22, 2002, pp. 6674–6681.

[37] M. Su et al., "The anti-angiogenic effect and novel mechanisms of action of combretastatin a-4," Sci Rep, vol. 6, Jun 2016, pp. 28 139–28 139, 27338725[pmid].

[38] R. Akizuki, S. Shimobaba, T. Matsunaga, S. Endo, and A. Ikari, "Claudin-5, -7, and -18 suppress proliferation mediated by inhibition of phosphorylation of akt in human lung squamous cell carcinoma," Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, vol. 1864, no. 2, 2017, pp. 293 – 302.

[39] Y. Ke-Da et al., "Genetic variants in gstm3 gene within gstm4-gstm2-gstm1-gstm5-gstm3 cluster influence breast cancer susceptibility depending on gstm1," Breast Cancer Research and Treatment, vol. 121, 2009, pp. 485–496.

[40] J. Zou et al., "Kinesin family deregulation coordinated by bromodomain protein ancca and histone methyltransferase mll for breast cancer cell growth, survival, and tamoxifen resistance," Mol Cancer Res, vol. 12, no. 4, Apr 2014, pp. 539–549, 24391143[pmid].

[41] A. Ruhul et al., "Nuclear pore complex protein, nup210 is a novel mediator of metastasis in breast cancer," NIH Research festival, 2018.

[42] J. Xie et al., "Knockdown of rab7a suppresses the proliferation, migration and xenograft tumor growth of breast cancer cells," Bioscience Reports, 2018.

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4

[43] K. Watanabe et al., "A novel somatic mutation of sin3a detected in breast cancer by whole-exome sequencing enhances cell proliferation through era expression," Scientific Reports, vol. 8, no. 1, 2018, p. 16000.

[44] Y. Wang, H. Xu, B. Zhu, Z. Qiu, and Z. Lin, "Systematic identification of the key candidate genes in breast cancer stroma," Cell Mol Biol Lett, vol. 23, Sep 2018, pp. 44–44, 30237810[pmid].

[45] H. A. M. Sakil, M. Stantic, J. Wolfsberger, S. E. Brage, J. Hansson, and M. T. Wilhelm, "Dnp73 regulates the expression of the multidrug-resistance genes abcb1 and abcb5 in breast cancer and melanoma cells - a short report," Cell Oncol (Dordr), vol. 40, no. 6, 2017, pp. 631–638, 28677036[pmid].

[46] X. Liu et al., "Discovery of microarray-identified genes associated with ovarian cancer progression," International journal of oncology, vol. 46, 04 2015.

[47] K. Xu, J. Cui, V. Olman, Q. Yang, D. Puett, and Y. Xu, "A comparative analysis of gene-expression data of multiple cancer types," PLoS One, vol. 5, no. 10, Oct 2010, pp. e13 696–e13 696, 21060876[pmid].

[48] E. Bailey et al., "Pulmonary vasculopathy associated with figf gene mutation," Am J Pathol, vol. 187, no. 1, Jan 2017, pp. 25–32, 27846380[pmid].

[49] F. Zhang et al., "Identification of key transcription factors associated with lung squamous cell carcinoma," Med Sci Monit, vol. 23, Jan 2017, pp. 172–206, 28081052[pmid].

[50] W. B. Kinlaw, P. W. Baures, L. E. Lupien, W. L. Davis, and N. B. Kuemmerle, "Fatty acids and breast cancer: Make them on site or have them delivered," J Cell Physiol, vol. 231, no. 10, Oct 2016, pp. 2128–2141, 26844415[pmid].

[51] F. Andriani et al., "Diagnostic role of circulating extracellular matrix-related proteins in non-small cell lung cancer," BMC Cancer, vol. 18, no. 1, Sep 2018, pp. 899–899, pMC6145327[pmcid].

[52] I. Cheuk et al., "Association of ep2 receptor and slc19a3 in regulating breast cancer metastasis," Am J Cancer Res, vol. 5, no. 11, Oct 2015, pp. 3389–3399, 26807319[pmid].

[53] O. Kowalczuk, J. Laudanski, W. Laudanski, W. E. Niklinska, M. Kozlowski, and J. Niklinski, "Lymphatics-associated genes are downregulated at transcription level in non-small cell lung cancer," Oncol Lett, vol. 15, no. 5, May 2018, pp. 6752–6762, 29849784[pmid].

[54] M. Bashir, S. Damineni, G. Mukherjee, and P. Kondaiah, "Activin-a signaling promotes epithelial-mesenchymal transition, invasion, and metastatic growth of breast cancer," Npj Breast Cancer, vol. 1, Aug 2015, pp. 15 007 EP –, article.

[55] H. Thomsen et al., "Inbreeding and homozygosity in breast cancer survival," Scientific Reports, vol. 5, Nov 2015, pp. 16 467 EP –, article.

[56] C. McCarty et al., "Alcohol, genetics and risk of breast cancer in the prostate, lung, colorectal and ovarian (plco) cancer screening trial," Breast Cancer Res Treat, vol. 133, no. 2, Jun 2012, pp. 785–792, 22331481[pmid].

[57] D. Mohelnikova et al., "The role of ABC transporters in progression and clinical outcome of colorectal cancer," Mutagenesis, vol. 27, no. 2, 03 2012, pp. 187–196.

[58] B. Weigelt, J. L. Peterse, and L. J. van't Veer, "Breast cancer metastasis: markers and models," Nature Reviews Cancer, vol. 5, Aug 2005, pp. 591 EP –, review Article.

[59] H. Kennecke et al., "Metastatic behavior of breast cancer subtypes," Journal of Clinical Oncology, vol. 28, no. 20, 2010, pp. 3271–3277, pMID: 20498394.

[60] J. Yin, W. Ren, X. Huang, J. Deng, T. Li, and Y. Yin, "Potential mechanisms connecting purine metabolism and cancer therapy," Front Immunol, vol. 9, Jul 2018, pp. 1697–1697, 30105018[pmid].

[61] C. Strell and F. Entschladen, "Extravasation of leukocytes in comparison to tumor cells," Cell Commun Signal, vol. 6, Dec 2008, pp. 10–10, 19055814[pmid].

[62] L. Edwards, R. Gupta, and F. V. Filipp, "Hypermutation of dpyd deregulates pyrimidine metabolism and promotes malignant progression," Mol Cancer Res, vol. 14, no. 2, Feb 2016, pp. 196–206, 26609109[pmid].

[63] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch, "Map kinase signalling pathways in cancer," Oncogene, vol. 26, May 2007, pp. 3279 EP –, review.

[64] G. W. McLean, N. O. Carragher, E. Avizienyte, J. Evans, V. G. Brunton, and M. C. Frame, "The role of focal-adhesion kinase in cancer – a new therapeutic opportunity," Nature Reviews Cancer, vol. 5, Jul 2005, pp. 505 EP –, review Article.

[65] Z. He, F. Tang, Z. Lu, Y. Huang, H. Lei, Z. Li, and G. Zeng, "Analysis of differentially expressed genes, clinical value and biological pathways in prostate cancer," Am J Transl Res, vol. 10, no. 5, May 2018, pp. 1444–1456.

Copyright (c) IARIA, 2020. ISBN: 978-1-61208-792-4