A machine learning integrated bioinformatics analysis for tissue specific breast cancer gene classification

[1]Ghazala Sultan, [2,*]Swaleha Zubair

[1,2]Department of Computer Science, Aligarh Muslim University, Aligarh, India.

Abstract: Machine learning techniques has been extensively utilized at early stages of biomedical research to analyze large datasets. This study aimed to develop machine learning models with strong prediction power and interpretability for gene classification between normal and cancer samples based on their expression level in different origins of tis-sues. We collected various candidate features from the clinical features of samples and generated filtered relatable features from original features set. Best features were selected through feature evaluation for classification of cancer specific genes. We used 30% of the data as a test dataset and 70% cases of data as a training and validation dataset on 7110 features from epithelial and stromal tissue. To develop the cancer gene prediction model, we considered five ma-chine learning algorithms: Logistic Regression, random forest (RF), support vector machine (SVM) and k-nearest neighbor (KNN) and C5.0. We found that random forest model shows the best learning model that produces the highest validation accuracy. In the random forest model, the classification accuracy of 95%, sensitivity is 0.926, specificity is 0.915, and AUC is 0.970. The developed prediction models show high accuracy, sensitivity, specificity and AUC in classifying among cancerous and healthy samples. This model could be used to predict BRCA in other patients with epithelial or stromal origin cancer. This study suggests that combination of multiple learning models may increase the cancer prediction accuracy.

Index Terms: Breast Cancer, Machine Learning, Supervised Learning, Unsupervised Clustering, Gene Classification

I. INTRODUCTION

Breast cancer is no longer viewed as a single disease, but rather a compilation of several distinct subtypes and often differentiated based on their location of origin [1]. Epithelial and Stromal are two types of origin of BRCA; the epithelial carcinomas are cancers that arise from the epithelial component that consists of the cells that line the lobules and terminal ducts [2]. Stromal carcinomas may originates from one or multiple stromal components including adipocytes, pre-adipocytes, fibroblasts, blood vessels, inflammatory cells, and ECM. In normal condition, these components are responsible for regulation throughout the developmental cycle [3]. Epithelial tissue is also the most common site for the development cancers. Carcinomas arise from epithelial tissue and account for as many as 90 percent of all human cancers. The most common cancers in humans occur in breast and colonic epithelium [4].

Over the past decade, machine learning techniques has received increasing attention in bioinformatics especially for cancer classification and candidate gene identification [5]. Machine learning methods for insight analysis in cancer classification has brought noticeable improvements in cancer diagnosis [6]. These algorithms help extensively in identification of candidate genes or predictive disease biomarkers in high-throughput sequencing datasets [7, 8]. Therefore, there is need for development of efficient classification methods to differentiate cancer specific gene and normally expressed genes with altered expression [9]. This approach may contribute efficiently for early detection of cancer.

This study is an attempt to develop machine learning models with strong prediction power and interpretability for gene classification be-tween normal and cancer samples based on their expression level in different origins of tissues. To develop the cancer gene prediction model, we considered five machine learning algorithms: ElasticNet, random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN) and C5.0. We finalized the final model as the best learning model that produces the highest validation accuracy based on classification accuracy percentage, sensitivity, and specificity and AUC scores.

II. DATA RETRIEVAL FOR META-ANALYSIS

The breast cancer dataset was retrieved from Gene Expression Omni-bus database of National center for Biotechnology Information (NCBI) with the accession ID GSE10797 and subjected for preprocessing. The data used in this study consist of total RNA was isolated from epithelial and stromal cells captured from normal breast tissue (n=10) and invasive breast cancer (n=56). The respective gene expression profiling was measured using microarray Affymetrix U133A 2.0 GeneChip. In total, the dataset encompasses 66 subjects including 5 control epithelial cells, 28 cancerous epithelial samples, 5 normal stromal samples and 28 cancerous stromal samples.

III. BIOINFORMATICS ANALYSIS

1) DATA PREPROCESSING AND SIGNIFICANT GENE MINING

The probe intensities from curated samples were subjected for normalization in R(v4.2.1) using Robust Multi-array Average (RMA). RMA for microarray data performs background correction, normalization and summarizes the probe level information [10]. To find the significant genes, we used Limma algorithm-based t-test (LAT).These differentially expressed genes were further explored for their expression levels in normal epithelial (NEp) versus cancerous epithelial (CEp), normal stromal (NSt) versus cancerous stromal (CSt) and cancerous epithelial (CEp) versus cancerous stromal (CSt).Furthermore, Benjamin and Hochberg False Discovery Rate and statistical t-test was used to calcu-late FDR and p-values to filter the significant genes. In order to identify DEGs (Differentially Expressed Genes), we considered a criteria $p < 0.05$ and $|\log FC| \geq 1$. A volcano plot and venn diagram were constructed, which represents p-value and fold change distribution and DEGs. A volcano plot dis-plays statistical significance (-log10 P value) versus magnitude of change (log2 fold change) and is useful for visualizing differentially expressed genes. The genes with $|\log FC| \geq$ were considered as upregulated and genes with $|\log FC| \leq 1$ were considered as downregulated genes.

2) Protein–Protein Interaction (PPI) Network

STRING was used for PPI networking of all DEGs [11]. The STRING database includes total 18,838 human proteins, with 25,914,693 interactions. The interactions in the Cytoscape v3. 4 have been used to analyze using various integral attributes. Protein interactions were initially uploaded into the Cytoscape v 3.4 and were assessed using integrated functions [12]. The highest score of confidence interaction was 0.9 in this analysis.

3) Enrichment Analysis of Gene Set

Both DEGs and modules function and pathway enrichment analysis was performed. The modules DEGs and hub genes were uploaded and parameters including mode-function, gene ID, species, molecular function, cellular component, biological process and active pathways were set to analyze the functional annotation of the identified candidate genes [13, 14]. The two-sided hypergeometric test (Benjamin-Hochberg method) with kappa score 0.96 and cutoff-value > 0.05 was performed for enrichment calculation.

IV. Machine Learning Based Gene Classification

For early detection and accurate classification of cancers, several ma-chine learning algorithms have been applied to microarray datasets, including Logistic regression, support vector machines (SVMs) [15], random forest [16], k-Nearest Neighbor [17] and C5.0 [18]. However, previous studies have shown that among popular techniques for multi-category classification of gene expression profiling datasets, SVMs have a dominant role, significantly outperforming all other methods.

The differently expressed gene out of total genes were treated as feature set to classify the normal and cancerous samples. Machine learning classifiers, including logistic regression, kNN, SVM (with the radial basis function used as the kernel function), RF and C5.0 were constructed and compared with cross validation of 10 folds. The best performing classifier was selected based on the prediction accuracy and confusion matrix. Additionally, the final classifier was built to identify the optimal set of cancer specific genes. Thereafter, the final classifier was examined and validated by using several other datasets. **Fig. 1** represent the workflow for cancer gene classification using machine learning classification techniques supported by bioinformatics analysis results.

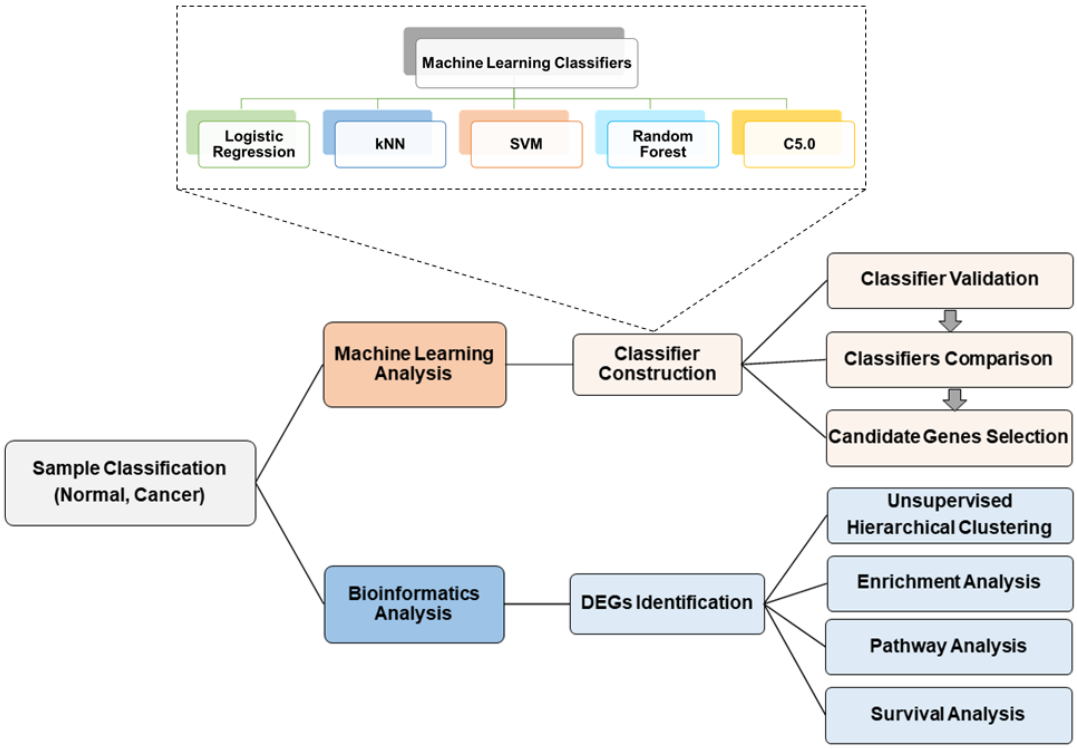

Figure 1. Schematic Diagram for the workflow of the study.

V. Results

The differential expression analysis identified 12 common genes which are expressed in both epithelial and stromal tissues, while 14 genes are uniquely expressed in epithelial breast tissue and 88 gene s were stromal specific. Fig. 2 shows that there were no significant gene found that were unique to normal epithelial and normal stromal samples. However, interestingly we identified there were differentially expressed genes which were specific to epithelial and stromal tissue.

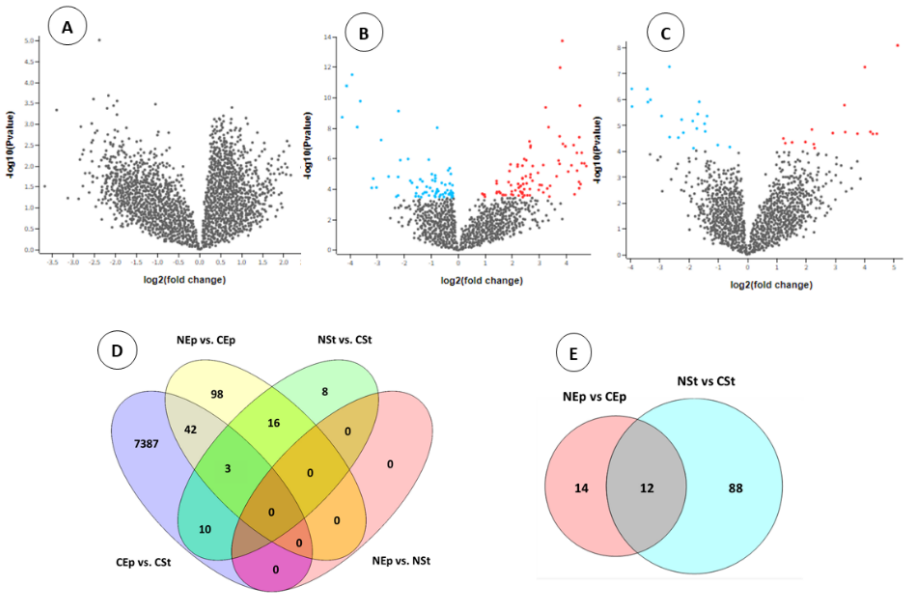

Figure. 2 Volcano plot for the transcriptomes of epithelium and stroma tissues (Padj<0.05). The plot shows the genes with distinct expression labelled by colors (blue - downregulated genes, red- upregulated genes). Distribution for NEp vs. NSt, (A), the distribution for CEp vs. NEp (B), the distribution for CSt vs. NSt (C). D and E shows the number of differentially expressed genes.

Table 1. Significant genes differentially expressed only in epithelial tissue.

Gene	Gene name	log2FC
Upregulated Gene		
SULF1	sulfatase 1	-2.671
FAM186A	family with sequence similarity 186 member A	-2.366
TLL2	tolloid like 2	-2.254
MMP1	matrix metallopeptidase 1	-1.874
COL5A1	collagen type V alpha 1 chain	-1.848
SNX10	sorting nexin 10	-1.704
MATN3	matrilin 3	-1.652
EPB41L5	erythrocyte membrane protein band 4.1 like 5	-1.447
RGS16	regulator of G-protein signaling 16	-1.019
CPM	carboxypeptidase M	-0.606
Downregulated Gene		
IGHA1	immunoglobulin heavy constant alpha 2	5.125
TFAP2C	transcription factor AP-2 gamma	2.287
RPL13	ribosomal protein L13	2.262
CCND3	cyclin D3	1.519
RPL9	ribosomal protein L9	1.223

Table 2. Significant genes differentially expressed only in stromal tissue.

Gene	Gene name	log2FC

Upregulated Gene		
CXCL2	C-X-C motif chemokine ligand 2	5.539
GABRP	gamma-aminobutyric acid type A receptor pi subunit	4.664
NTRK2	neurotrophic receptor tyrosine kinase 2	4.567
PROM1	prominin 1	4.524
SYNM	synemin	4.476
KRT14	keratin 14	4.473
LTF	lactotransferrin	4.467
OPRPN	opiorphinprepropeptide	4.302
RGS2	regulator of G-protein signaling 2	4.269
ALDH1A3	aldehyde dehydrogenase 1 family member A3	4.186
IGHA2	immunoglobulin heavy constant alpha 2	4.068
Downregulated Gene		
COL1A2	collagen type I alpha 2 chain	-3.201
COL3A1	collagen type III alpha 1 chain	-3.027
NAT1	N-acetyltransferase 1	-2.854
GPRC5A	G protein-coupled receptor class C group 5 member A	-2.145

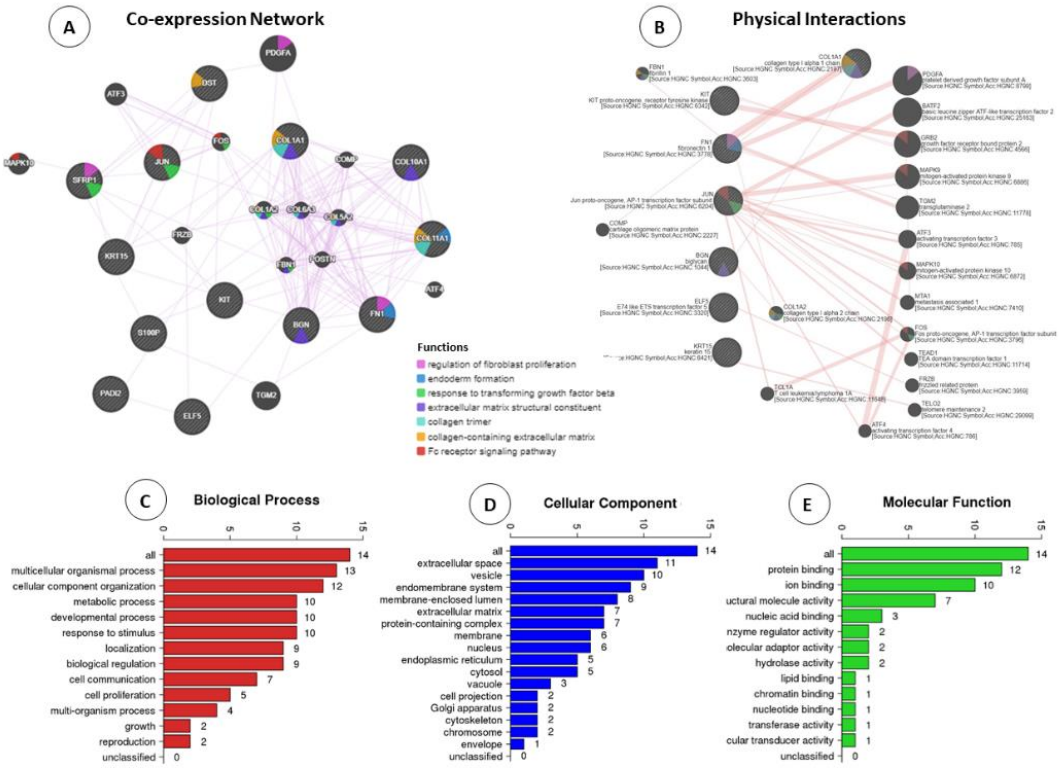

Figure 3. Gene interaction network for epithelium and stroma tissues; co-expressed genes (A) and physically interacted genes (B). The enriched biological processes (C), cellular components (D) and molecular functions (E) are represented based on their enrichment score.

To identify cancer related genes, the network nodes were ranked based on similarity with breast cancer driver genes. We constructed a profile for training set using over-represented terms based on GO terms, gene expression, semantics, pathways, phenotypes and scores network node based on similarity with training set. For this study, driver genes were used as training set and network genes were used as test set. All training parameters were enabled, false discovery rate (FDR) was applied. The p-value cut-off of 0.05 was set,d minimum count was set as 2 and 1500 random sampling were enabled. The 7,110 differently expressed gene out of total 21,270 expressed genes were treated as feature set to classify the normal and cancerous samples. Fig 4. shows the gene importance selected by the ML classifiers.

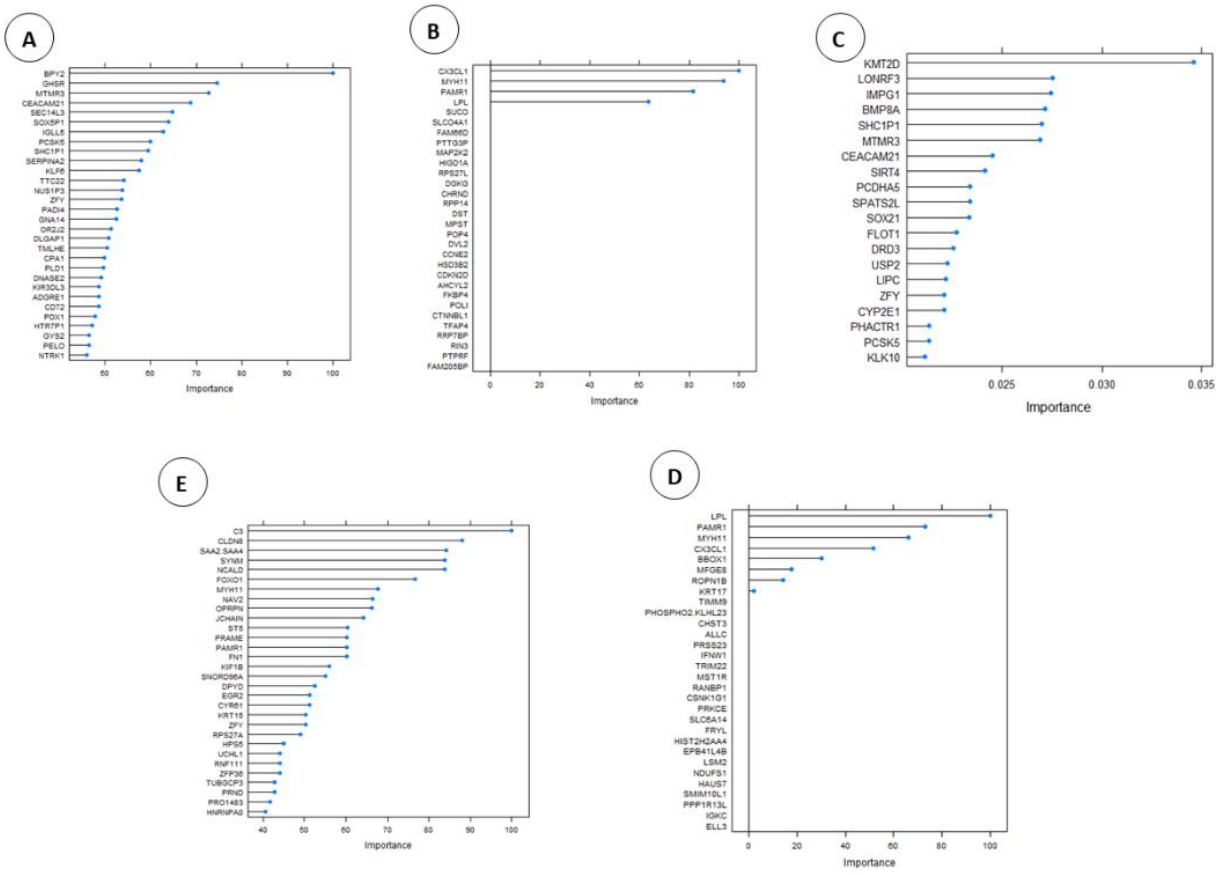

Figure 4. Gene importance selected by the ML classifiers (A) ElasticNet, (B) KNN, (C) SVM, (D) Random Forest and (E) C5.0.

Table 3. The accuracy obtained from five machine learning classifiers.

ML model	Accuracy	
	Epithelial	Stromal
ElasticNet	0.95 (alpha=0, lambda=0.2)	0.93 (alpha=0, lambda=0.2)
KNN	0.90 (k=9)	0.97 (k=9)
SVM	0.883 (sigma=8.949e-05, C=0.25)	0.831 (sigma=8.31e-05, C=0.75)
Random Forest	0.9467 (mtry=2)	0.96 (mtry=2)
C5.0	0.883 (trials =1)	0.871(trials=1, model=rules)

VI. Conclusion

This study is an attempt to classify cancer specific genes in epithelial and stromal breast cell utilizing bioinformatics analysis integrated with machine learning classification approach. The 30% of the data was split as a test dataset and 70% of the cases as a training set on 7,110 genes from epithelial and stromal tissue. The cancer gene prediction model was designed with the machine learning classifiers including ElasticNet, random forest (RF), support vector machine (SVM) and k-nearest neighbor (KNN) and C5.0. From the considered classification models, the random forest model came out as the best learning model which classifies tissue specific genes with highest accuracy in comparison to other models (ElasticNet, SVM, KNN and C5.0). In the random forest model, the classification accuracy of 94.67% for epithelial tissue and 96% accuracy for stromal tissue. The developed prediction models show high accuracy, sensitivity, specificity and AUC in classifying among cancerous and healthy samples for both types of tissues. The implemented model could be used to predict breast cancer in the patients with epithelial or stromal origin cancer. This study suggests that combination of multiple learning models may increase the cancer gene prediction accuracy.

References

1. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T. C., & Ren, G. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes & diseases, 5(2), 77–106 (2018).

2. McCuaig, R., Wu, F., Dunn, J., Rao, S., and Dahlstrom, J. E. The biological and clinical significance of stromal-epithelial interactions in breast cancer. Pathology, 49(2), 133–140 (2017).
3. Wiseman, B. S., &Werb, Z..Stromal effects on mammary gland development and breast cancer. Science (New York, N.Y.), 296(5570), 1046–1049 (2002).
4. Mazlan, A.U.; Sahabudin, N.A.; Remli, M.A.; Ismail, N.S.N.; Mohamad, M.S.; Nies, H.W.; Abd Warif, N.B. A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. Processes, 9, 1466 (2021).
5. Hinck, L., & Näthke, I. Changes in cell and tissue organization in cancer of the breast and colon. Current opinion in cell biology, 26, 87–95 (2014).
6. Gupta S, Gupta MK, Shabaz M and Sharma A. Deep learning techniques for cancer classification using microarray gene expression data. Front. Physiol. 13:952709 (2022).
7. Zhang X, Jonassen I, Goksøyr A. Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In: Helder I. N, editor. Bioinformatics. Brisbane (AU): Exon Publications; Chapter 4. (2021)
8. Koppad, S., Basava, A., Nash, K., Gkoutos, G. V., &Acharjee, A. Machine Learning-Based Identification of Colon Cancer Candidate Diagnostics Genes. Biology, 11(3), 365. (2022)
9. Roy, S., Kumar, R., Mittal, V. et al. Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning. Sci Rep 10, 4113 (2020).
10. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England), 4(2), 249–264 (2003).
11. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research, 49(D1), D605–D612 (2021).
12. Chen Y-C, Hsiao C-C, Chen K-D, Hung Y-C, Wu C-Y, Lie C-H, Liu S-F, Sung M-T, Chen C-J, Wang T-Y. Peripheral lung cancer patients treated with first line combination chemotherapy. PLoS ONE One 8(2):e57053 (2013).
13. Sultan, G., Zubair, S, Tayubi, I. A., Dahms, H. U. and Madar, I H. Towards the early detection of ductal carcinoma (a common type of breast cancer) using biomarkers linked to the PPAR(γ) signaling pathway. Bioinformation, pp. 15(11):799-805 (2019),
14. Sultan, G., Zubair, S., Madar, I. H. and Anandaram H. Computational Approach to Identify Regulatory Biomarkers in the Pathogenesis of Breast Carcinoma. International Journal of Advanced Computer Science and Applications, Vol. 13, No. 6, (2022).
15. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer genomics & proteomics, 15(1), 41–51 (2018).
16. Toth, R., Schiffmann, H., Hube-Magg, C., Büscheck, F., Höflmayer, D., Weidemann, S., Lebok, P., Fraune, C., Minner, S., Schlomm, T., Sauter, G., Plass, C., Assenov, Y., Simon, R., Meiners, J., & Gerhäuser, C. Random forest-based modelling to detect biomarkers for prostate cancer progression. Clinical epigenetics, 11(1), 148 (2019).
17. Ehsani, R., & Drabløs, F. Robust Distance Measures for kNN Classification of Cancer Data. Cancer informatics, 19, 1176935120965542 (2020).
18. Duan, S., Cao, H., Liu, H., Miao, L., Wang, J., Zhou, X., Wang, W., Hu, P., Qu, L., & Wu, Y. Development of a machine learning-based multimode diagnosis system for lung cancer. Aging, 12(10), 9840–9854 (2020).