VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2021, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 9, Number 1, January-March 2021 *pp: 24-29*

Identifying Key Genes of Liver Cancer by Using Random Forest Classification

Muhammad Sohaib Roomi[1, *], Muhammad Adeel Ashraf[1], Muhammad SOHAIB AKRAM [2]

[1] Department of Computer Sciences, *University of Management and Technology, Lahore,Pakistan*
[2]*Govt Graduate College Mailsi, Pakistan*
Corresponding Author: F2018279051@umt.edu.pk

ABSTRACT
Liver cancer is considered as one of the most deadly cancer. To devise a treatment which is helpful to eradicate, it is inevitable to identify potential biomarkers which are very important in the development of liver cancer. To identify the pathways and key genes we use different enrichment analysis techniques such as pathway analysis and functional analysis. To identify biomarkers we constructed a network which is named as protein protein interaction network to analyse by selecting different network nodes. Our results show that we identified those biomarkers like ESR1 and TOP2 successfully which are potential biomarkers for liver cancer. In addition to that our method can be applied to other different datasets which are for different diseases to choose key genes.

KEYWORDS
Liver Cancer, Key genes, Biomarkers.

JOURNAL INFO
HISTORY: Received: February 15, 2021 Accepted: March 15, 2021 Published:..March 30, 2021

1. INTRODUCTION

When we talk about most dangerous cancers, liver cancer is up there in the list. It is rated 5th most common cancers in the world. And person with this type of cancer usually has very low survival rate.

One of the main reason is lack of effective treatment and pathways. That is why liver cancer is still a hot topic to discuss. Identifying biomarkers of this disease accurately. We combine different datasets to compute differentially expressed genes with high accuracy and if we individually perform operations on each dataset then not more accurate results can be achieved than the results which are obtained by networking of genes datasets.

With this proposed paradigm14 very important key genes are observed during PPI mechanism in which 12 biomarkers proved to be highly significant for the progression of this disease while their 2 had very high relation with the others[1].

Different genes like TOP2 and ESR1 proved to be a biomarkers for liver cancer moreover these key genes are also used in other studies[2].

To understand liver cancer, vast research has been made which results in creation of very high number of key genes datasets of liver cancer and easy availability of these datasets provide an opportunity to study it more closely. And creation of these large amount of datasets resulted in the creation of different tools.

For instance different machine learning tools are developed that are used to closely analyse the sequence of genes and proteins like DNA and RNA[3].

Results showed that this paradigm can identify accurate key genes due to which this method can be easily adjusted with other biological datasets such as DNA, single polymorphism and RNA sequencing[4].

To find the treatment of this cancer it's very important to know it's mechanism that how it evolves itself to the level where it is almost incurable.

And to understand it's mechanism it is very important to know genes which has direct effect in its progression. Once we are able to identify these key genes then it can be cured by targeting those specific genes.

To understand liver cancer, vast research has been made which results in creation of very high number of key genes datasets of liver cancer and easy availability of these datasets provide an opportunity to study it more closely. creation of these large amount of datasets resulted in the creation of different tools.

For instance different machine learning tools are developed that are used to closely analyse the sequence of genes and proteins like DNA and RNA. Networking of these genes datasets provides an effective analysis of the mechanism which results in the creation of different networks. For instance Long et al. developed Protein -Protein Network from differentially expressed genes to identify key biomarkers from microarray data. It has been unarguably recognized that by networking different datasets can provide a better tool to identify different biological methods. Lisette also combined different networks and gene expressions to identify genetic drivers of disease. Guelzim has devised a structure of yeast network[4]. Kim developed a structural mechanism to differentiate normal genes and poorly modulated in different diseases. Vanunu designed a computational model which is known as PRINCE, which is used to prioritize defective genes and complex protein associations[5].

This work is licensed under a Creative Commons Attribution 3.0 License.

2. MATERIAL AND METHODS.

More than 400 cases data for Liver cancer was obtained by GDC which contain and 3096 mutations. These mutations of liver cancer correlated to give 575 genes in total. Due to different sample size and different size in effects makes it difficult to compare. Then we follow two layered graph approach[6]. Reason for not using single layered graph approach is that this provides very less amount of information about tumor and gene sample prediction. It is necessary to observe information of relation of genes between them and because of this there must be another method to map this, which is another source of information that helps to improve the accuracy of prediction model. It is also observed from previous studies that using multilayer graphs yields more accuracy than single graph[7].With this proposed system,14 very important biomarkers are recognized during PPI mechanism in which 12 biomarkers are considered of very high significance in the development of disease while their 2 had very high relation with the others[8]. Different genes like TOP2 and ESR1 proved to be a biomarkers for liver cancer moreover these key genes are also used in other studies.

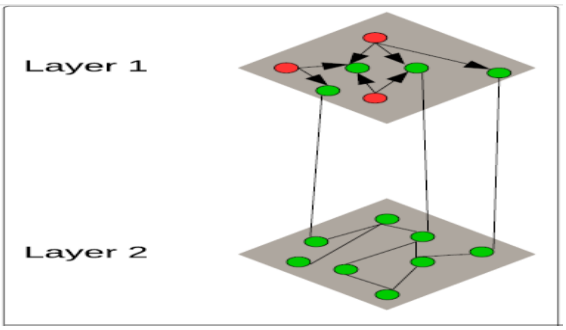


Figure-1: Graphical illustration of network that represents genes in green and tumor defective genes in red.

We analysed protein protein interaction network that has weights for homosapiens. Usually there are 8 channels tha tare linked to protein protein interaction network[9]. We used two techniques to transform interaction network into gene network:

i) Mapped protein names to according to encoding genes.
ii) We selected the maximum weight that connects any protein.

Table 1 illustrates the summarization of network is observed during PPI mechanism model[10][11].

Table-1: Brief Summarization of double layer method

Property	Value
Number of tumor samples	4086
Number of genes	575
Number of relations between tumor samples and genes (hasGene)	2252
Co-occurrence	1166
Co-expression	8470
Neighborhood	1892
Text mining	2883
Combined	3586
Physical	839

3. SOLUTION APPROACH.

Our main purpose is to make prediction link in tumor samples and biomarkers so we use graph as an input. We devised an elastic net regression with formulation to observe the role of every gene which helps in liver cancer progression by Sparse Coefficient that also helps to select the biomarkers. Correlation of genes mean they will reside in one group. To get grouping effect the model should be rigorously convex

$$O_1 = \|y - \beta X\|^2 + \lambda\left(\frac{(1-\alpha)}{2}||\beta||^2 + \alpha||\beta||\right) \quad (1)$$

Different genes which interact with each other in the protein protein network normally has same functions. Furthermore, to examine the working of liver cancer different protein protein network are formed using differentially expressed genes. Then those interactions were chosen to design a network with a threshold of .04.

Different genes which interact with each other in the protein protein network normally has same functions. Furthermore, to examine the working of liver cancer different protein protein network are formed using differentially expressed genes. Then those interactions were chosen to design a network with a threshold of .04. represents up regulation and three down regulation.

Usually different network functions form different communities. The interaction between key genes and differentially expressed genes can have high impact in the network. Research of communities in the genes is very complex work[12]. Therefore Girvan Neman is an algorithm that is best fit for this method. The main work of this algorithm is calculate betweenness among the edges. This betweenness measure the edges that are underlying in the communities. Following are the steps that are used in this process:

1. Calculate the score of every edge.
2. Choose the edges with high value and eliminate these edges from the system.
3. Re-evaluate the betweenness value from the leftover nodes.

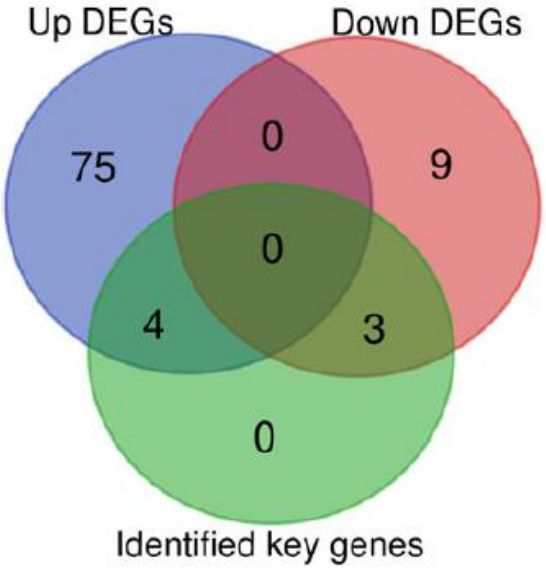


Figure-2: The Diagram of recognized biomarkers. Four

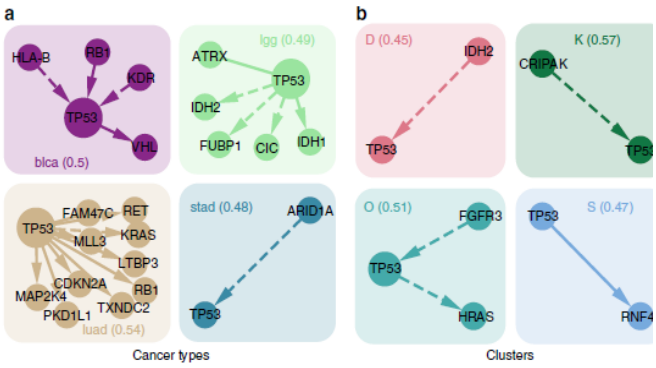


Figure-3: Look TP53 closely with the interactions of other genes in **a** Cancer and **b** [13][14]

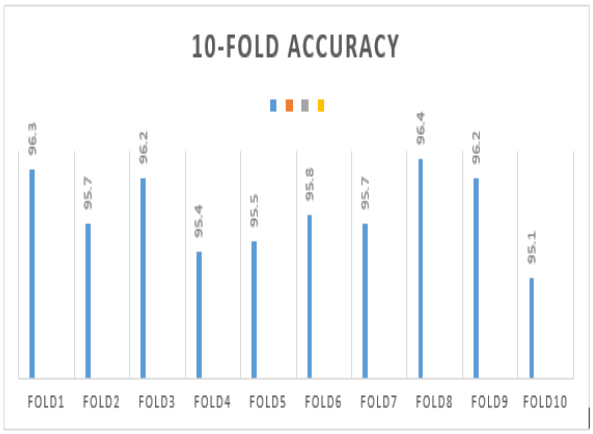
Moreover this model do require such high accurate and clear results for which we need to use training sets carefully. Those results which are very absurd and containing error are left alone and it selects those which are close to required results. The algorithm was designed by Glay to implement it on the network. This algorithm of Glay is used to partition the protein protein network. To identify the key genes more closely the protein interaction network of different type of biomarkers should be analysed to get the attributes of nodes[15]. Normally key nodes have high effect on the other edges in system. Betweenness is map of network which resides on the smallest path. In this graph of distance vertex 'v' betweenness can be expressed as[16]:

$$C_B(v) = \sum_{i,j:i\neq j,i\neq v,j\neq v} \frac{g_{ivj}}{g_{ij}}, \qquad (2)$$

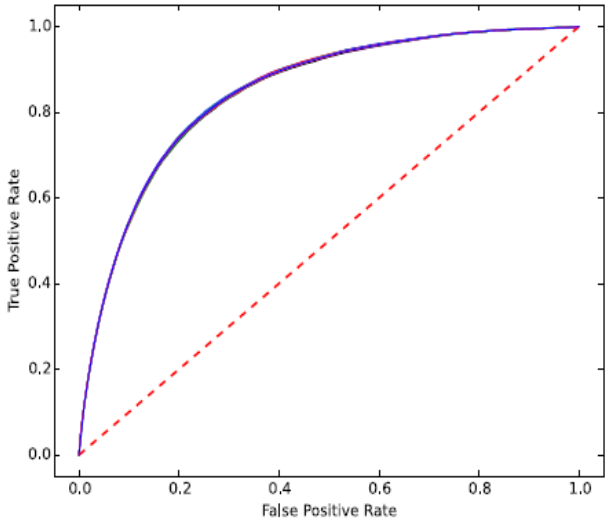
Where 'g', 'j', 'k' represents the shortest path from 'i' to 'j' in 'k'. To measure the influence of nodes globally, the network which we named as similarity network was changed into distance network.

4. RESULTS AND DISCUSSIONS

During training of our model, it was necessary to do it the best way we can. The training must be in the way that it has enough instances for training. There should be enough instances, if this requirement does not meet then the training of model would not be ideal and training would not be in the proper way. Accuracy is very important when it comes to classification[17]. While working with the limited datasets it is recommended to use K fold cross validation and use the value of K big. On the other hand while dealing with large datasets it is recommended to use leave one out technique [21-44]. We use Random Forest Algorithm as classifier. It is an ensemble algorithm. It means it performs classification by combining other algorithms for classifying objects. For example using combination of Naïve Bayes, Decision tree and SVM and then choosing result for an object class. So we perform K-Fold cross validation and predict the accuracy of system. When we check results so often then surely the quality of results and approximate segments will improve and that is the purpose of testing it. But we cannot check it after every invocation because then it cannot be called training. According to this we make predictions of those inputs and then comparison of actual results and predicted results is made and decided that how much accuracy is achieved. The clustering is very much based on the data mutation, so we make clusters' comparison regardless of any clinical information. Some differences in clusters can be described by the profiles of mutations. Clusters are highly important for prediction of survival. From this experiment we also observe the true positive and false positive rate which is very important for the prediction of model that identifies key genes for liver cancer called training. According to this we make predictions of those inputs and then comparison of actual results and predicted results is made and decided that how much accuracy is achieved. The clustering is very much based on the data mutation, so we make clusters' comparison regardless of any clinical information. Some differences in clusters can be described by the profiles of mutations. Clusters are highly important for prediction of survival. From this experiment we also observe the true positive and false positive rate which is very important for the prediction of model that identifies key genes for liver cancer.



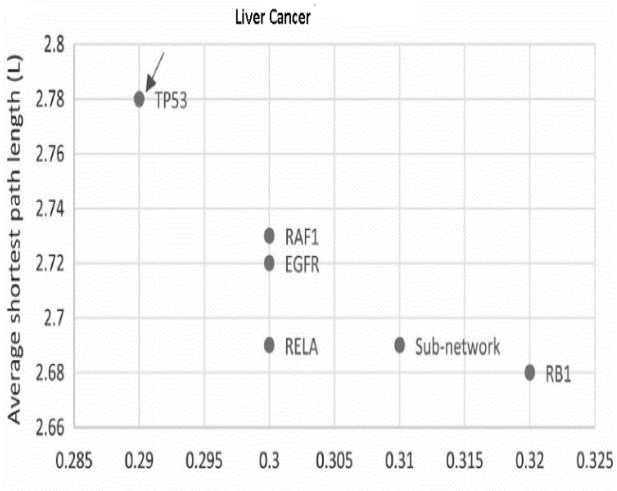
Graph-1: Illustration of K-Fold Accuracy in each fold[11]



Graph-2: Represents the curved in which red line represents guess taken randomly while blue line is mean score of K fold [10]

Table 2: shows the Comparative Analysis of RFA (our) and PPI (Paper) techniques

Methods	Accu	Sen	Spec	Mcc
RFA (Random Forest)	95.72	91.3	97.5	86.73
PPI network (Networking of Multiple Datasets)	93.87	89.4	95.2	85.86



Graph-2: Average shortest path graph legth

5. CONCLUSION

In this paper, the proposed system which forecasts different biomarkers and key genes of disease which have very high importance for the development of this type of cancer. These identified key genes can be termed as the important biomarkers of disease and can be helpful for the treatment of disease. To analyse the potential biomarkers of Liver cancer we collected liver cancer data from national cancer institute and analysed their differentially expressed genes. Using the classifier, some of key genes were selected and termed as key genes of liver cancer. This has been demonstrated that these key genes have player significant role in the progression of disease. Out of these genes two have very high interaction with the potential biomarkers. By combination of these, it has been demonstrated that our model predict these key genes effectively. The most important thing is our model can be adapted for the other data like RNA-Sequence to choose key genes of disease

REFERENCES

[1] A. Fujimoto et al., "Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer," Nat. Genet., vol. 48, no. 5, p. 500, 2016.

[2] J. Zhan, Y. Cai, S. He, L. Wang, and Z. Yang, "Tandem Molecular Self-Assembly in Liver Cancer Cells," Angew. Chemie Int. Ed., vol. 57, no. 7, pp. 1813–1816, 2018.

[3] S. M. Inavolu et al., "IODNE: An integrated optimization method for identifying the deregulated subnetwork for precision medicine in cancer," CPT pharmacometrics Syst. Pharmacol., vol. 6, no. 3, pp. 168–176, 2017.

[4] S.-P. Deng and W.-L. Guo, "Identifying key genes of liver cancer by networking of multiple data sets,"

IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 16, no. 3, pp. 792–800, 2018

[5] J. Kuipers et al., "Mutational interactions define novel cancer subgroups," Nat. Commun., vol. 9, no. 1, pp. 1–10, 2018.

[6] A. Keliris, H. Salehghaffari, B. Cairl, P. Krishnamurthy, M. Maniatakos, and F. Khorrami, "Machine learning-based defense against process-aware attacks on industrial control systems," in 2016 IEEE International Test Conference (ITC), 2016, pp. 1–10.

[7] P. Maji and E. Shah, "Significance and functional similarity for identification of disease genes," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 6, pp. 1419–1433, 2016.

[8] H. Güney and H. Öztoprak, "Microarray-based cancer diagnosis: repeated cross-validation-based ensemble feature selection," Electron. Lett., vol. 54, no. 5, pp. 272–274, 2018.

[9] J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," IEEE Access, vol. 7, pp. 4232–4238, 2018.

[10] J. Li, W. Dong, and D. Meng, "Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 15, no. 6, pp. 2028–2038, 2017.

[11] L. T. T. Scaria and T. Christopher, "A Bio-inspired Algorithm based Multi-class Classification Scheme for Microarray Gene Data," J. Med. Syst., vol. 43, no. 7, p. 208, 2019.

[12] M. Timilsina, H. Yang, R. Sahay, and D. Rebholz-Schuhmann, "Predicting links between tumor samples and genes using 2-Layered graph based diffusion approach," BMC Bioinformatics, vol. 20, no. 1, p. 462, 2019.

[13] Z. Xu, Y. Zhou, Y. Cao, T. L. A. Dinh, J. Wan, and M. Zhao, "Identification of candidate biomarkers and analysis of prognostic values in ovarian cancer by integrated bioinformatics analysis," Med. Oncol., vol. 33, no. 11, p. 130, 2016.

[14] W. Du, K. Dickinson, C. A. Johnson, and L. N. Saligan, "Identifying Genes to Predict Cancer Radiotherapy-Related Fatigue with Machine-Learning Methods," in Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, p. 527.

[15] S.-P. Deng, L. Zhu, and D.-S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 13, no. 1, pp. 27–35, 2015.

[16] L. Zhang, H. Liu, Y. Huang, X. Wang, Y. Chen, and J. Meng, "Cancer progression prediction using gene interaction regularized elastic net," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 14, no. 1, pp. 145–154, 2017.

[17] B. Liang, C. Li, and J. Zhao, "Identification of key pathways and genes in colorectal cancer using bioinformatics analysis," Med. Oncol., vol. 33, no. 10, p. 111, 2016.

[18] H. Q. Pham, L. Rueda, and A. Ngom, "Predicting Breast Cancer Outcome under Different Treatments by Feature Selection Approaches," in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2017, p. 617

[19] A. Amala and I. A. Emerson, "Identification of target genes in cancer diseases using protein–protein interaction networks," Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 8, no. 1, p. 2, 2019.

[20] H. Liu, Y. Zhao, L. Zhang, and X. Chen, "Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal," Mol. Ther. Acids, vol. 13, pp. 303–311, 2018.

[21] Saeed, S.; Mahmood, M. K.; Khan, Y. D., An exposition of facial expression recognition techniques. Neural Computing and Applications 2018, 29 (9), 425-443.

[22] Butt, A. H.; Khan, Y. D., CanLect-Pred: A cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences. IEEE Access 2019, 8, 9520-9531.

[23] Amanat, S.; Ashraf, A.; Hussain, W.; Rasool, N.; Khan, Y. D., Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. Current Bioinformatics 2020, 15 (5), 396-407.

[24] Ilyas, S., Hussain, W., Ashraf, A., Khan, Y. D., Khan, S. A., & Chou, K. C. (2019). iMethylK-PseAAC: Improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou's 5-steps rule. Current Genomics, 20(4), 275-292.

[25] Hussain, W.; Rasool, N.; Khan, Y. D., A Sequence-Based Predictor of Zika Virus Proteins Developed by Integration of PseAAC and Statistical Moments. Combinatorial chemistry & high throughput screening 2020, 23 (8), 797-804.

[26] Khan, Y. D.; Alzahrani, E.; Alghamdi, W.; Ullah, M. Z., Sequence-based Identification of Allergen Proteins Developed by Integration of PseAAC and Statistical Moments via 5-Step Rule. Current Bioinformatics 2020, 15 (9), 1046-1055.

[27] Mahmood, M. K.; Ehsan, A.; Khan, Y. D.; Chou, K.-C., iHyd-LysSite (EPSV): Identifying Hydroxylysine Sites in Protein Using Statistical Formulation by Extracting Enhanced Position and Sequence Variant Feature Technique. Current Genomics 2020, 21 (7), 536-545.

[28] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., IPhosS (Deep)-PseAAC: Identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps rule. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2020.

[29] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. Current Bioinformatics 2020, 15 (8), 937-948.

[30] Shah, A. A.; Khan, Y. D., Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. Scientific Reports 2020, 10 (1), 1-10.

[31] Awais, M.; Hussain, W.; Rasool, N.; Khan, Y. D., iTSP-PseAAC: Identifying Tumor Suppressor Proteins by Using Fully Connected Neural Network and PseAAC. Current Bioinformatics 2021, 16 (5), 700-709.

[32] Hussain, W.; Rasool, N.; Khan, Y. D., Insights into Machine Learning-based approaches for Virtual Screening in Drug Discovery: Existing strategies and streamlining through FP-CADD. Current Drug Discovery Technologies 2021, 18 (4), 463-472.

[33] Khan, Y. D.; Khan, N. S.; Naseer, S.; Butt, A. H., iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. PeerJ 2021, 9, e11581.

[34] Malebary, S. J.; Khan, R.; Khan, Y. D., ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins. IEEE Access 2021, 9, 68788-68797.

[35] Malebary, S. J.; Khan, Y. D., Evaluating machine learning methodologies for identification of cancer driver genes. Scientific reports 2021, 11 (1), 1-13.

[36] Malebary, S. J.; Khan, Y. D., Identification of Antimicrobial Peptides Using Chou's 5 Step Rule. CMC-COMPUTERS MATERIALS & CONTINUA 2021, 67 (3), 2863-2881.

[37] Naseer, S.; Ali, R. F.; Khan, Y. D.; Dominic, P., iGluK-Deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. Journal of Biomolecular Structure and Dynamics 2021, 1-14.

[38] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., NPalmitoylDeep-PseAAC: A Predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. Current Bioinformatics 2021, 16 (2), 294-305.

[39] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. Analytical Biochemistry 2021, 615, 114069.

[40] Khanum, S., Ashraf, M. A., Karim, A., Shoaib, B., Khan, M. A., Naqvi, R. A., ... & Alswaitti, M. Gly-LysPred: Identification of Lysine Glycation Sites in Protein Using Position Relative Features and Statistical Moments via Chou's 5 Step Rule.

[41] Lv, H., Dao, F. Y., Zhang, D., Yang, H., & Lin, H. (2021). Advances in mapping the epigenetic modifications of 5-methylcytosine

[42] (5mC), N6-methyladenine (6mA), and N4-methylcytosine (4mC). Biotechnology and Bioengineering.

[43] Zulfiqar, H., Sun, Z. J., Huang, Q. L., Yuan, S. S., Lv, H., Dao, F. Y., ... & Li, Y. W. (2021). Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods.

[44] Liu, Y., Wang, X., & Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Briefings in bioinformatics, 20(1), 330-346.

[45] Zhang, D., Xu, Z. C., Su, W., Yang, Y. H., Lv, H., Yang, H., & Lin, H. (2021). iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics, 37(2), 171-177.